Investeşte în oameni!
FONDUL SOCIAL EUROPEAN
Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007 – 2013
Axa prioritară: 1 „Educaţia şi formarea profesională în sprijinul creşterii economice şi dezvoltării societăţii bazate pe cunoaştere”
Domeniul major de intervenţie: 1.5 „Programe doctorale şi postdoctorale în sprijinul cercetării”
Titlul proiectului: Proiect de dezvoltare a studiilor de doctorat în tehnologii avansate- ”PRODOC”
Numarul de identificare al contractului: POSDRU 6/1.5/S/5
Beneficiar: Universitatea Tehnică din Cluj-Napoca

ing. **Adriana Cornelia STAN**

TEZA DE DOCTORAT

ROMANIAN HMM-BASED TEXT-TO-SPEECH SYNTHESIS

WITH INTERACTIVE INTONATION OPTIMISATION

SINTEZA TEXT-VORBIRE ÎN LIMBA ROMÂNĂ BAZATĂ PE MODELE MARKOV ŞI OPTIMIZAREA INTERACTIVĂ A INTONAȚIEI

Conducător ştiinţific
Prof.dr.ing. **Mircea GIURGIU**

UNIVERSITATEA TEHNICÃ
DIN CLUJ-NAPOCA

FACULTATEA DE ELECTRONICÃ, TELECOMUNICATII ŞI TEHNOLOGIA INFORMATIEI

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).

Contents

List of Abbreviations

ASR	Automatic Speech Recognition
CART	Classification and Regression Trees
CDF	Cerevoice Development Framework
DCT	Discrete Cosine Transform
F0	fundamental frequency or pitch
GV	Global Variance
HMM	Hidden Markov Model
HSMM	Hidden Semi-Markov Model
HTS	HMM-Based Text-to-speech Synthesis
IPA	International Phonetic Alphabet
MDL	Minimum Description Length
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MLSA	Mel Log Spectrum Approximation
MOP	Maximal Onset Principle
MSD-HMM	Multi Space Distribution Hidden Markov Model

PSOLA	Pitch Synchronous Overlap and Add
SAMPA	Speech Assessment Methods Phonetic Alphabet
TTS	Text-to-speech

List of Tables

List of Figures

Chapter 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.

1.3 Thesis Outline

The thesis is organised as follows:

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.

Chapter 2

Background

2.1 Text-to-speech Synthesis

Text-to-speech (TTS) synthesis is a method for deriving human-like speech from a text input. Fig. 2.1 shows the basic blocks of a TTS system. The process can be more easily understood if a parallel to learning a new language is made. Given a text in a new language, the first step is to determine the text segments which have to be preprocessed for a correct reading, such as numbers, abbreviations, neologisms and so on. Then, each letter has to be transposed in an acoustic correspondent or phoneme. Individual correspondents are not enough, as context influences the sound of a given letter. Having the correct succession of phonemes can then be concatenated into syllables, words, phrases and so on. Simple phrasing, duration and basic intonation are then assigned. And, at last the physical process of speech production, through mechanical articulation of the sounds takes place. If the person is a more advanced speaker of that language, emphasis and prosody are more likely to be reproduced correctly, similar to a native speaker.

In the same way, text-to-speech systems evolved from the simple reproduction of individual sounds through wooden tubes, to state-of-the-art speech synthesisers which use advanced semantical analysis and can output high-quality expressive speech. The entire system is usually broken down into two major components: *text processing* and *speech synthesis*. Each of them imply extensive analysis and synthesis methods with their correspondent arising problems.

The **goals** of a TTS system according to [Taylor, 2009] are to clearly get the message

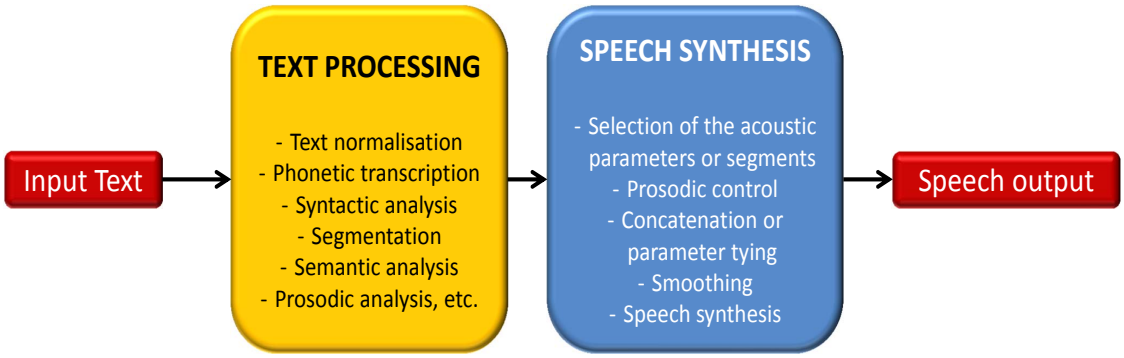


Figure 2.1: Block diagram of a text-to-speech system

across to the listener in terms of **intelligibility** and **naturalness**, and to be able to synthesise **any given input text**. This means that the text processor has to be able to transform any input text into a sequence of labels, and that the speech synthesiser has the means of outputting qualitative speech from any sequence of input labels.

The need for a text-to-speech system can be emphasised through its **applications**. The initial purpose of TTS was for visually impaired people to have access to written texts without the help of the Braille alphabet. With the appearance of analogue and digital storing devices, the speech synthesisers were used in even more applications. By concatenating pre-recorded speech segments, the system could output a limited number of combinations between the samples. This type of synthesiser is still used in client-information applications, such as automated answering and GPS machines or ATMs. More advanced TTS systems are used in intelligent dialogue applications or in combination with automatic speech recognition, even translating applications from one language to another.

The following sections give an overview of the standard synthesis methods and the use of prosody within the text-to-speech systems. The last section presents the Romanian synthesisers available.

2.2 Speech Synthesis Methods

Speech production is a complex process which involves a large number of computational resources and memory. Aside from the even more complex task of caring out a dialogue, even the reading aloud of a text implies training and processing on behalf of a person. Over the years multiple methods of speech synthesis have been proposed. One of the earliest proofs of the so-called *talking heads* are mentioned for Aurrilac (1003 A.D.), Albert Magnus (1198-1280) or Roger Bacon (1214-1294). In 1779, the Russian researcher Christian Kratzenstein, created models of the human vocal tract which could reproduce the *a,e,i,o* and *u* vowels [Giurgiu and Peev, 2006]. The first electronic synthesiser was the VODER (Voice Operation DEmonstratoR) created by Homer Dudley at Bell Laboratories in 1939 [Dudley, 1940]. The VODER produced only two basic sounds: a tone generated by a radio valve to produce the vocal sounds and a hissing noise produced by a gas discharge tube to create the sibilants. These basic sounds were passed through a set of filters and an amplifier that mixed and modulated them into the resulted speech. To get the machine to actually speak required an operator to manipulate a set of keys and a foot pedal to convert the hisses and tones into vowels, consonants, stops, and inflections. Fortunately, nowadays, speech synthesisers have evolved to a point where their intelligibility and naturalness is comparable to human speakers and their operation requires a minimum amount of training on behalf of the speaker.

Based on the main method of generating the speech signal, speech synthesisers can be classified into **rule-based** and **corpus-based**. In rule-based methods no pre-recorded speech samples are used, each sound is defined by a fixed set of parameters. Corpus-based methods use either the speech samples or segments of them, or derive their parameters from the direct analysis of the speech corpus. Some argue that corpus-based,especially when using speech samples is not a true synthesis method, because the signal is not synthesised from scratch, although it is the most commonly used one. In this sense a different definition of the speech synthesis notion can be given [Taylor, 2009], i.e. *the output of a spoken utterance from a resource in which it has not been prior spoken.*

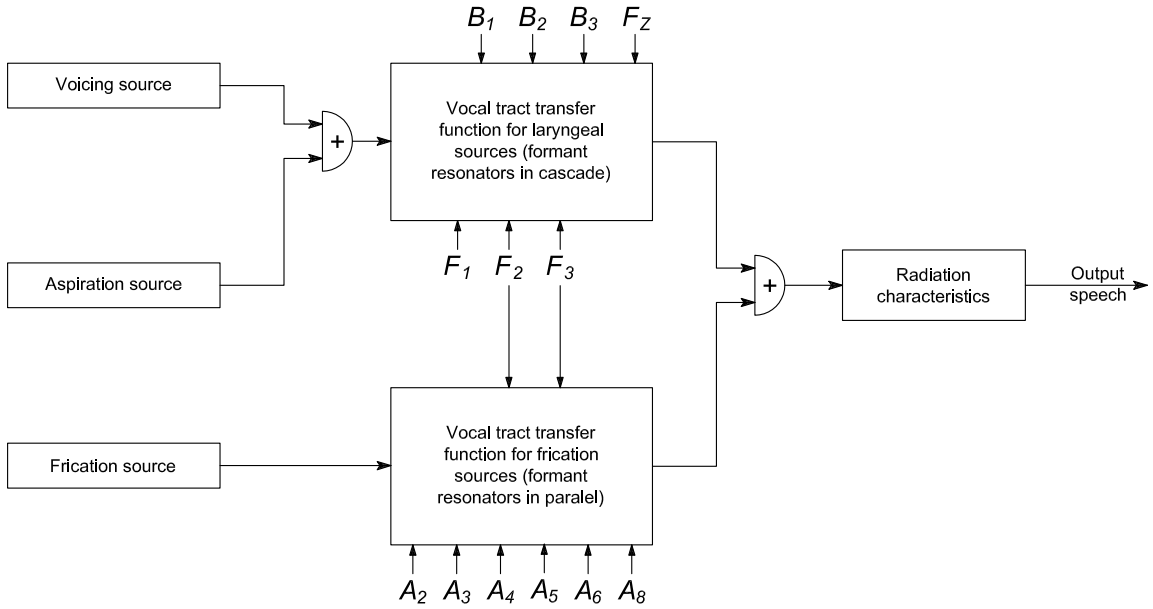


Figure 2.2: Klaat's formant synthesiser (after [Benesty et al., 2007])

2.2.1 Rule-based

Formant synthesis

Formant synthesis determines a set of rules on how to modify pitch, formants frequencies and other parameters from one sound to the other [Huang et al., 2001]. It is based on the source-filter model of speech production. In formant synthesis, the formant resonances are represented by a number of filters having as input a train of impulses for the voiced segments and white noise for the unvoiced segments.

The most representative model of formant synthesis is the one described by [Klatt, 1980], which later evolved into the commercial system of MITalk [Allen et al., 1987]. There are around 40 parameters which describe the formants and their respective bandwidths, and also a series of frequencies for nasals or glottal resonators. A parallel structure of second order FIR filters is implemented for the fricatives and stops, and a cascade structure for the voiced sounds [Allen et al., 1987]. Fig. 2.2 presents a simplified version of the Klatt synthesiser structure.

The problem with the formant synthesis is that the source-filter model itself has the drawback of not including the reaction of the filter unto the source. Another drawback is

the fact that the acoustic realisation of a sound varies over time, and cannot be represented by a fixed set of parameters. The same speaker asked to repeat the same word multiple times, will use different duration and intonations. Therefore, the formant synthesis lacks the modelling of the minute variations that make a long duration speech sample natural.

Prosody in formant synthesis can be achieved by modifying the set of frequencies or filter parameters. However, this implies an extended study of the prosodic effects on pitch and formants.In [Wolf, 1981] or [Apopei and Jitcă 2007] certain rules for prosody control are derived, but their results cannot be generalised, because of the particular characteristics of parameter analysis and synthesis used.

Articulatory synthesis

Articulatory synthesis has the potential of becoming one of the best synthesis methods. It uses mechanical and acoustic models of speech production to synthesise speech [Benesty et al., 2007]. The physiological effects are modelled, such as the movement of the tongue, lips, jaw, and the dynamics of the vocal tract and glottis. Biomechanical, aerodynamic and acoustic studies are also involved. For example [Bickley et al., 1997] uses lip opening, glottal area, opening of nasal cavities,constriction of tongue, and rate between expansion and contraction of the vocal tract along with the first 4 formant frequencies. This is a method which articulatory synthesis with the formant-based one, thus trying to alleviate the drawbacks of the later.

Unfortunately, the model is very complex and there is still a lack of analysis methods of the processes involved. A complete articulatory model would also include the electric impulses of the nerves and muscle movement of the entire phonatory apparatus. The use of magnetic resonance imaging has offered some more elaborate models of muscle movement and thus the results of the articulatory synthesis have improved.

However, the results of this type of synthesis are still far from being considered natural because of the use of partially heuristic determined rules and the fact that the acoustic processes vary from speaker to speaker. The physical characteristics of a person, such as length of vocal tract and tongue size influence the mechanical movement in speech production. Accurate physiological understanding of speech production is also lacking and thus the parameters of the model cannot be fully determined.

Prosody is not yet an issue for this type of synthesis because of the early stages of development in which it still is.

2.2.2 Corpus-based

Concatenative synthesis

Concatenative synthesis is the most commonly used method in commercial systems. It became a popular choice once the storage and computational characteristics of the digital devices became more and more advanced. The basic idea is the use of pre-recorded speech samples of fixed or variable lengths, which can fully capture the fine details of speech. This aspect was not possible in the rule-based methods.

In this type of method, an utterance is synthesised by concatenating several natural segments of speech (Fig. 2.3). The samples are stored in a database, indexed by the phonetic content along with prosodic markers, context or other additional information. Samples of speech can include utterances, words, syllables, diphones or phonemes. Based on the type of segment stored in the database, the concatenative synthesis is either **fixed inventory** – segments in the database have the same length, or **unit selection** – segments have variable length and the system makes a decision of the best match.

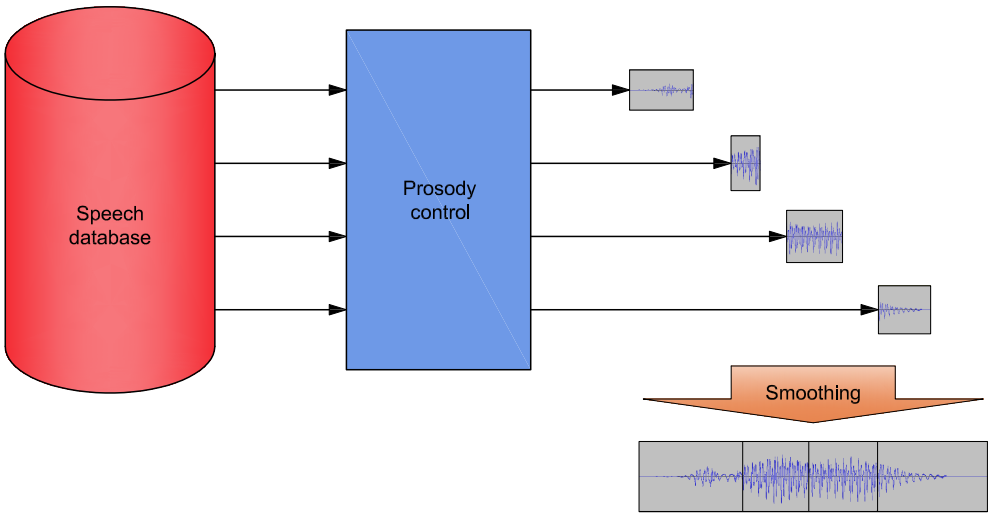


Figure 2.3: Basic principle of a concatenative unit selection speech synthesis system (after [Benesty et al., 2007])

The most common fixed inventory concatenative synthesis is the **diphone** concatenation [Black et al., 1999, Lambert and Breen, 2004]. A diphone in this case is defined from the middle of the first phoneme to the middle of the second one. Using this type of segmentation avoids the concatenation discontinuities at phoneme boundaries. For a simple diphone concatenation system, the database or speech corpus would include a single repetition of all the diphones in a language. Some more elaborate systems use diphones in different context (e.g. beginning, middle or end of word) and with different prosodic events (e.g. accent, variable durations etc.). Two major problems with this approach are: the coarticulation[1] [Olive et al., 1993] effects over longer units – which cannot be captured by the diphones; and the concatenation errors – diphones taken from different contexts have different amplitudes and pitch values.

Another type of fixed inventory system is based on the use of **syllables** as the concatenation unit [Saito et al., 1996, Matoušek et al., 2005, Buza, 2010], but although it can reduce some of the concatenation discontinuities[2], the speech database is hard to design. The average number of syllables in one language is the order of thousands.

The best concatenative synthesis solution is **unit selection** [Black and Campbell, 1995], [Hunt and Black, 1996], which uses variable length speech samples. The samples are selected using scores to determine the best match, and can be phonemes, diphones, syllables, words, or even entire phrases. The speech corpus design is minimum, although extended databases provide better results. Coarticulation problems are solved in unit selection by introducing the target cost (Eq. 2.1) and concatenation costs (Eq. 2.2) [Black and Campbell, 1995]. Target cost represents the cost of selecting a particular unit from the database, while concatenation cost is the cost of using that unit in the utterance. The best unit is selected using a Viterbi-like search algorithm over the two cost functions.

$$C_{target}(u_i, u_s) = \sum_{j=1}^{N} w_t^{(t)} c_j^{(t)}(u_i, u_s), \tag{2.1}$$

u_i represents the candidate unit, and u_s the current unit.

[1] A phoneme's acoustic production depends on the context in which it is present.

[2] There are some theories which state that the basic unit of speech is the syllable and the coarticulation effects between them is minimum

$$C_{concatenation}(u_{i-1}, u_i) = \sum_{k=1}^{M} w_k^{(c)} c_k^{(c)}(u_{i-1}, u_i), \tag{2.2}$$

u_i represents the newly selected unit, and u_{i-1} the previously selected unit.

Prosody in concatenative synthesis was initially achieved by the use of extended speech corpora (i.e. tens of hours of speech) which included different prosodic realisations of the same speech unit. Later, with the use of PSOLA (Pitch Synchronous Overlap and Add) [Moulines and Charpentier, 1990] and other more advanced techniques, the prosody could be modelled for each unit individually, at the expense of some loss in naturalness. However, the best results are still obtained with large scale corpora. Prosodic manipulation would be easier to achieve if the waveform would have a parametrised form, such as the case of the next synthesis method.

Although this type of method is applied in most of the best commercial, there are some major disadvantages of the concatenative synthesis. First of all the need for an extended speech corpus is not practical, as it requires hours of recording, segmentation and annotation. Second of all, prosody control from within the corpus is hard to achieve, because the segments have to be correctly hand labeled and concatenated. An important problem relies in the flexibility of the system: once a system is built, changing the speaker requires re-recording the database, annotating it and adjusting the parameters to the specific speaker.

Statistical Parametric Synthesis

Unit selection synthesis, although providing one the of best quality synthetic speech lacks the flexibility of the output speech. Quality is directly determined by the speech corpus and the unit selection algorithms. Parametrising the speech waveform is a solution to the generalisation problem of the synthetic speech in concatenative systems. Parametric corpus-based synthesis implies the use of a pre-recorded speech corpus from which it extracts a selection of parameters. Thus speech synthesis becomes a statistical analysis of a speech corpus. Parameters are clustered according to context and prosodic features.

The most important parametric technique is the one based on hidden Markov models

(HMMs), a concept borrowed from automatic speech synthesis and with very good applicability and flexibility within speech synthesis as well. A first attempt to model speech using HMMs is that of [Falaschi et al., 1989], but the results were unnatural and did not come to the attention of the specialists. With the introduction of the HMM-based Speech Synthesis System (HTS) [Zen et al., 2007b], some of the initial problems were solved, and this method became the choice for research in speech synthesis. In HTS, speech is modelled through a 3 state HMM for each phoneme. Each state includes mel frequency cepstral coefficients and F0 with their delta and delta-delta features, and state duration (Fig. 2.4). Decision trees are employed for the context clustering of the feature vectors to ensure no low or zero occupancy states. Contextual factors include phonetic, accentual and syntactic features.

From the target phoneme sequence a sentence of HMMs is derived using the Maximum Likelihood (ML) algorithm. Over smoothing of the spectral sequence is partially solved by the global variance (GV) principle [Toda and Tokuda, 2007], which maximises the dynamic variation of the speech parameters.

Different parameter sets are used in [Acero, 1999] and [Kawahara et al., 1999]. The [Acero, 1999] approach uses formants as acoustic observation, thus trying to overcome some of the formant synthesis problems. A very good method of parametrisation is that of [Kawahara et al., 1999], called STRAIGHT and which uses source and spectral parameters in the form of: a mixed-excitation model based on a weighted combination of fundamental frequency and noise, and a set of aperiodicity bands.

The advantages of the parametric synthesis refer to:

- the small footprint necessary to store speech information;
- automatic clustering of speech information– removes the problems of hand-written rules;
- even small training corpora can result in good quality of the synthetic speech, if the corpus is well designed;
- generalisable – even if for a certain phoneme context there is not enough training data, the model might be clustered along with similar parameter characteristics;
- flexibility – the trained models can be easily adapted to other speakers or voice characteristics with minimum amount of adaptation data.

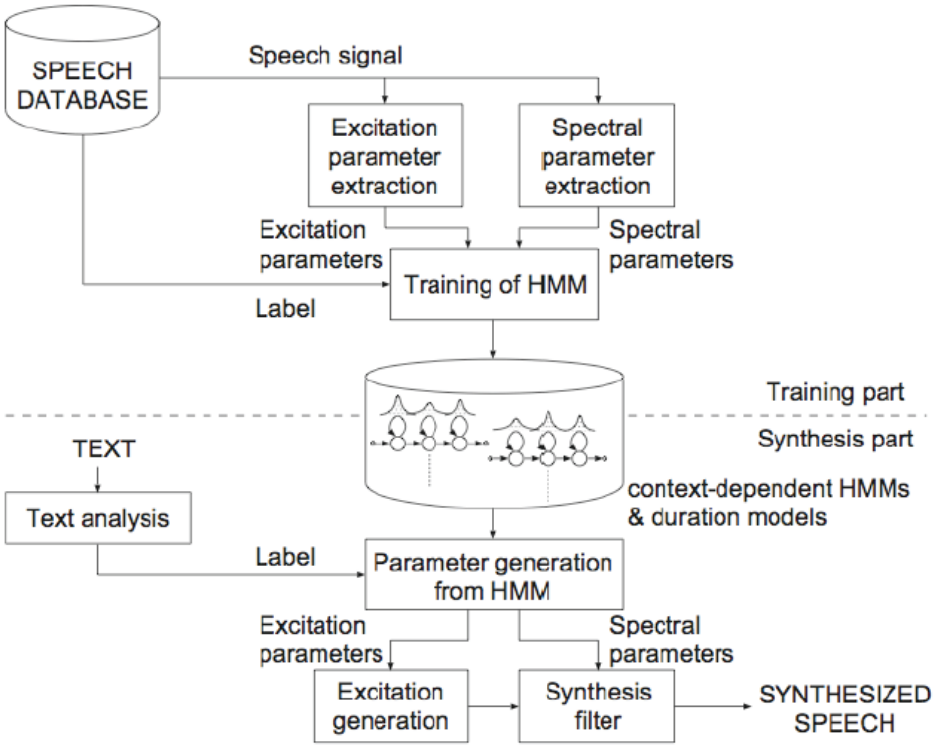


Figure 2.4: Common HMM-based speech synthesis system (after [Black et al., 2007])

.

And of course, there are also disadvantages such as low speaker similarity due to the use of a parametrisation method which cannot capture the fine details of speech. Training on a large database leads to high computational requirements during the training stage, but the synthesis part is still minimum computational consuming. Another disadvantage, but can be also considered as an advantage is the fact that the output is highly dependent on the parametrisation method, which can be modified and adapted according to new researches.

Prosody in HTS is achieved by modifying the F0 decision trees or the sequence of states generated in the synthesis stage. Because it is a parametric method which uses the ML principle, the modification of the F0 contour can be made, without affecting the spectral characteristics. Therefore, it is easy to test new F0 contours for the same utterance, without affecting the naturalness of the synthetic speech.

Detailed theoretical overview of the HMM model and HTS can be found in chapter 4.

2.3 Speech Synthesis Systems for Romanian

The Romanian speech synthesisers are not that numerous, and except for a few commercial systems, their quality is rather poor. Below, a list of the available Romanian text-to-speech systems to the best of the author's knowledge can be found. It can be observed that the techniques used in the research systems are outdated.

- **Ivona - Carmen** `www.ivona.com` - unit selection - commercial
- **Nuance -Simona** `www.nuance.com` - unit selection - commercial
- **Loquendo - Ioana** `www.loquendo.com` - unit selection - commercial
- **AT&T Bell Labs** - concatenation of diphones, context-sensitive allophonic units or even of triphones
- **MBROLA** `http://tcts.fpms.ac.be/synthesis/mbrola.html`, [Dutoit et al., 1996] - diphone concatenation with overlap and add algorithm for pitch and duration control
- **Romanian Formant Synthesis** [Jitcă et al., 2002] - formant synthesis, with several prosody control techniques defined for it, such as the ones presented in the works of [Jitcă et al., 2008], [Apopei and Jitcă 2005] or [Apopei and Jitcă 2007]
- **RomVox** [Ferencz, 1997] - LPC-based parametric - allows for manual adjustment of some prosodic parameters
- **RomSyn** [Giurgiu and Peev, 2006] - diphone concatenation
- **LIGHTVOX** [Buza, 2010] - syllable concatenation
- **BRVox** [Bodo, 2009] - diphone concatenation with simple intonation pattern assignment based on the research of [Hirst and Cristo, 1998]
- **Baum - Ancutza** `http://www.baum.ro/` -diphone concatenation - commercial
- **Phobos TTS** `http://www.phobos.ro/demos/tts/index.html` - based on MBROLA
- **eSpeak** `http://espeak.sourceforge.net/` - formant synthesis
- **LingvoSoft Talking Dictionary** `http://www.lingvosoft.com/` - unavailable for testing

More information on Romanian syntheis systems and prosodic modelling can be found in chapters 4 and 5.

2.4 Summary

This chapter has presented a theoretical overview of the main speech synthesis methods available, while emphasizing on their strengths and weaknesses. The rule-based methods use a set of rules for defining the appropriate parameters for the speech units, mainly phonemes. This type of method is hard to control, as the variability of the parameters is limited to the studied cases. Their naturalness is poor, but they can be well put to use in perception experiments and speech production evaluations. In this category, articulatory synthesis is a potential high-qualitative system, but the complexity of the models involved puts it on hold until more data can be analysed.

Corpus-based or data driven methods on the other hand are the choice of both commercial and research systems. They are high-quality synthesisers and require far less heuristically determined parameters. The main categories of this technique are concatenative synthesis and parametric synthesis. The concatenative appoach actually plays back pre-recorded speech samples in the order that the utterance requires. Parametric synthesisers are more flexible, and the method of parametrisation can be easily modified. The mainstream system for parametric synthesis is the HMM-based, which uses methods similar to those defined in automatic speech recognition, thus the adaptation of proposed methods in both fields can be interchangeable.

Prosody control aspects are presented for each of the methods underlining the advantages and disadvantages resulting from them. It is clear that prosody control is still a debatable subject in each of the methods, as a full correct prosodic model, has not yet been defined. HMM-based synthesis offers the easiest way of controlling the intonation without the need of modifying the rest of the parameters.

In the final section of this chapter, the available text-to-speech systems available for Romanian are enumerated. Best results are of course obtained by the commercial systems with unit selection methods. It can be seen that there is a lack of freely available high-quality synthesisers for research purposes. This concluding in the choice for one of the topics for this thesis, as an HMM-based Romanian speech synthesisers with all the resources made available and a full documentation of the method of implementation.

Chapter 3

Resource Development for a Romanian Parametric Speech Synthesiser

3.1 Introduction

Any text-to-speech system requires a language-dependent acquisition of resources in the training and development stages of both the text processor and the speech synthesis parts. Resources are essential for the understanding of the particular phonological phenomena and lead to a higher quality of the resulted system. The influence of the available resources for a specific language, can be stated as follows. **Text resources** influence the resulted speech through:

- correct phonetic transcription of the input text;
- valid accent positioning;
- text normalisation;
- correct phrasing and pause assignment;
- the possibility to synthesise various text styles;
- in some advanced speech synthesisers, the focus of the utterance is established by semantic analysis.

While the **speech resources** enhance the speech synthesis by:

- the possibility to synthesise any valid succession of letters/phonemes in the selected language;
- the correct estimation of the parametric models – for parametric synthesis;
- multiple prosodic models for the same phoneme– parametric synthesis;
- sufficient unit and prosodic choices – in concatenative unit selection;
- lack of concatenation errors – for concatenative synthesis.

In the context of resource availability, Romanian is considered to be an *under-resourced* language. Several research groups devote their efforts into building solid tools and resources for the study of the Romanian language. Unfortunately, most of the times, these elements are not visible, standardised or public.

For promoting Romanian speech technology research, especially in speech synthesis, it is therefore essential to improve the available infrastructure, including free large-scale speech databases and text-processing front-end modules. This thesis provides a set of freely available, web-advertised resources and demonstrations developed for the purpose of a Romanian text-to-speech system. The resources are not optimal, but as the results show, they can be successfully used for a parametric HMM-based TTS system.

3.2 Text Resources

The resources described in the following sections comprise the tools, resources and preprocessing of data that lead to the Romanian HTS labeller used in the speech synthesis system. Text processing is one of the most challenging aspects of any TTS system in a new language. The great variability among different language groups and local specific alterations to standard spelling or grammar make it an important and vital part of any TTS system.

For Romanian, there are a few projects and publications regarding text processing, such as [Frunză et al., 2005] or [Giurgiu and Peev, 2006]. [Frunză et al., 2005] is an adaptation of the BALIE system for Romanian, which includes language detection, tokenisation, sentence boundary detection and part-of-speech tagging. [Giurgiu and Peev, 2006] is an elaborate description of the building of a Romanian TTS system, RomSyn. It includes phonetic transcription and some intonation patterns derived directly from text.

For the purpose of this study, a new text processor was developed, based on the Cerevoice development framework (CDF) [Aylett and Pidcock, 2007]. Language-dependent data has been gathered and probabilistic models have been trained; the front-end outputs HTS format labels comprising 53 kinds of contexts [Zen et al., 2007a].

3.2.1 The Text Corpus

The first step in the development of the front-end of the speech synthesiser was the acquisition of a large text corpus used in the training of individual components. Between August 2009 and September 2009, 4506 newspaper articles containing over 1,700,000 words were trawled from the online newspaper "Adevărul".

Table 3.1: Top 40 most frequent words of the text corpus and their relative frequencies

Word	Frequency [%]	Word	Frequency [%]
de	5.31	am	0.44
şi	3.38	ce	0.40
a	3.23	al	0.39
în	2.82	mii	0.37
la	2.10	fi	0.35
să	1.35	va	0.32
o	1.35	sunt	0.32
cu	1.26	ca	0.30
din	1.24	dar	0.27
care	1.11	după	0.26
pe	1.03	lui	0.26
că	1	e	0.25
nu	0.97	iar	0.22
au	0.93	sau	0.21
pentru	0.91	vor	0.21
mai	0.87	ar	0.19
un	0.84	dacă	0.18
se	0.79	ne	0.18
fost	0.58	le	0.17
este	0.56	prin	0.17

Because of the lack of uniformity between the writing styles of different authors and of the use of HTML tags within the text, some preprocessing of the text had to be carried out. This included diacritic normalisation[1], correct spelling and exclusion of embedded tags for links and videos. The result was a collection of short newspaper articles – 15 rows on average.

From the entire text corpus the top 65,000 most frequent words have been selected. These comprise the **lexicon**, which has been later phonetically transcribed and the accent positioning has been inserted. The 65,000 words were checked against the DEX online database [DEX online-webpage, 2011]. This means that all of the words exist in Romanian, have a valid spelling and there are no proper names or neologisms within the lexicon. The 65,000 words represent 4% of the total number of words existent in the DEX database. The top 40 most frequent words with their relative frequencies are presented in Table 3.1. It can be observed that, as expected, the top 40 include mainly prepositions and common verbs, in accordance with [Vlad and Mitrea, 2002].

3.2.2 Phonetic Transcription

Phonetic transcription represents the key starting point for a text-to-speech system. It determines the correct pronunciation and has a direct impact on the speech corpus design. The first step is of course to establish the phonetic inventory. The Romanian phonetic inventory generally consists of 7 vowels, the short vowel *i*, 4 semivowels and 20 consonants. However, linguists extend this set of phonemes, by the inclusion of allophones and rare case exception pronunciations [Giurgiu and Peev, 2006].

Several phonetic transcribers have been presented in the works of [Domokos et al., 2011, Toma and Munteanu, 2009, Burileanu et al., 1999] and [Ordean et al., 2009]. The methods used are various rule-based or semi-statistical approaches, and take into account an extended list of phonemes and allophones for Romanian.

A big advantage in using HTS, is that through clustering some of the phonemes may determine classes correspondent to their allophones, thus eliminating the need of an extended phonetic vocabulary. This leads to the reasoning behind the approach presented in this thesis, which uses only 32 phonemes. Table 3.2 shows the phone set used in the

[1]In Romanian there are two standards for the *ş and* ţ letters, one with cedilla and one with comma.

Table 3.2: Phone set used for the phonetic transcription, in SAMPA notation.

vowel	a @ 1 e i o u i_0
semivowel	e_X j o_X w
nasal	m n
plosive	b d g k p t
affricate	ts tS dZ
fricative	f v s z S Z h
trill	r
approximant	l
silence/pause	*'sil' 'pau'*

experiments in a SAMPA[2] notation. The set of phonemes, presented also in Appendix A, does not include allophones and rare case exceptions, but rather a minimal set which suffices the needs of the TTS system.

Romanian is mainly a phonetic language and letter-to-sound rules are quite straightforward. However there are several exceptions, which occur mainly in vowel sequences, such as diphthongs and triphthongs. Therefore a lightly supervised automatic learning method for letter-to-sound rules was adopted, as follows: From the text corpus, the top 65,000 most frequent words were extracted. General simple initial letter-to-sound rules were written manually by a native speaker. These rules were used to phonetically transcribe the complete list of words. To deal with the exceptions above, the pronunciations of 1000 words chosen at random were checked, and corrected where necessary, by a native speaker. Using this partially-corrected dictionary of 65,000 words, letter-to-sound rules were automatically learnt using a classification and regression tree (CART) [Breiman et al., 1984]. The accuracy of the obtained model is about 87%, measured using 5-fold cross validation. A small additional lexicon was manually prepared to deal mainly with neologisms, whose pronunciations are typically hard to predict from spelling. This custom lexicon is first searched by the TTS system and letter-to-sound rules are applied afterwards. The complete list of letter-to-sound rules written in Festival format can be found in Appendix B. Some further particular rules were added in the lexicon manually.

[2]Speech Assessment Methods Phonetic Alphabet

Some examples of phonetic transcription using the phone set presented in Appendix A are the following:

inedite i_n_e_d_i_t_e

george dz_e_o@_r_dz_e

excursie e_k_s_k_u_r_s_i_e

foarte f_o@_a_r_t_e

3.2.3 Accent Positioning

Romanian has a stress accent, that generally falls on the rightmost syllable of a prosodic word[3]. While a lexically marked stress pattern with penultimate stress exists, any morphologically derived forms will continue to follow the unmarked pattern [Chitoran, 2002].

However, the online SQL database of the Romanian Explicative Dictionary (DEX: `http://dexonline.ro/`) provides accent positioning information. Using this information from DEX directly, an accent location dictionary for the 65,000 most frequent words in the text corpus was prepared. In the resulting TTS system, the same lightly supervised algorithm as in the case of phonetic transcription, is used for the accent positioning.

3.2.4 Syllabification Using the Maximal Onset Principle

Romanian syllabification has 7 basic rules, which apply to the orthographic form of the words. But these can be affected by morphology, such as compound words or hyphenated compounds.

The CDF uses for syllabification, the Maximal Onset Principle (MOP), which states that intervocalic consonants are maximally assigned to the onsets of syllables rather than the coda, in conformity with universal and language-specific conditions. MOP has been evaluated for most of the European languages and attains an average accuracy of over 70%.

The MOP has not been previously applied to Romanian. In order to use this principle, onset consonant groups and vowel nuclei have been defined. A partial evaluation was

[3]The root and derivational material, but excluding inflections and final inflectional vowels

performed on 500 random manually syllabified words.The MOP principle was applied and attained an accuracy of 75%.

One of the major exceptions occurs in the vowel-semivowel-vowel groups, where both the vowel-semivowel and semivowel-vowel group can be a diphthong, thus a nuclei. For example, the word *caiet* contains the vowel groups *a-i* and *i-e*, which can be both diphthongs. This affects the syllabification as the result of the MOP application leads to the syllabified form *cai-et*.

Another important exception is represented by the compound words, where the syllabification is based on morphological decomposition and not the standard rules. But this cannot be addresses in either the 7 basic rules, not the MOP, because it requires a higher level knowledge, including word decomposition or lexemes.

3.2.5 Part-of-Speech Tagging

Part-of-Speech (POS) tagging is mainly used for word disambiguation, phrasing or focus assignment. The latest methods include elaborate statistical methods and some artificial intelligence [Naseem et al., 2009]. For Romanian, [Tufis et al., 2008], [Frunză et al., 2005] and [Calacean and Nivre, 2009] describe preliminary results but their resources are not available. There is, however, a freely available online tool using an HMM-based POS tagger [Sabou et al., 2008]. The work has not been published in any scientific journal or conference, but the authors report in an internal evaluation that the accuracy of the POS tagger is around 70%.

Using this tool, the entire text corpus was split into sentences and POS tagged. No additional evaluation has been performed and thus, only two categories have been used in the output of the text processor: **feature** – which includes nouns, verbs, adjectives and some adverbs – and **content** – the rest of the words.

3.2.6 The Lexicon

As a side development of the text resources, the phonetic transcription and accent positioning have been gathered in a so-called **lexicon**. The list of top 65,000 most frequent words in the text corpus have been phonetically transcribed using the list of phonemes presented in Appendix A. Along with the phonetic transcription, the accent position was

inserted for each word, using the DEX online database [DEX online-webpage, 2011]. An example entry in the lexicon is shown below, and an extended extract from it can be found in Appendix C.

abandoneze a0_b_a0_n_d_o0_n_e1_z_e0

All the vowels and semivowels have an accent marker attributed, with "0" and "1" as possible values. "1" marks the accent. There can be only one accent in each word.

3.3 Speech Resources

As well as text resources, speech resources for Romanian are scarce. There are some limited speech corpora, such as [Teodorescu et al., 2010] which is a small collection of the Romanian sounds (vowels, consonants, diphthongs and triphthongs) and a few sentences read with different emotions and by different speakers. In [Kabir and Giurgiu, 2010] the development of the Romanian version of the GRID corpus [Cooke et al., 2006]. Some several other small speech corpora are presented in [Giurgiu and Peev, 2006], [Bodo, 2009] and [Ferencz, 1997]. All of the databases have been built for a particular purpose and cannot be properly applied to other systems.

For any text-to-speech system, the speech resource is the key feature. If for the text processing some aspects can be left aside, such as normalisation – presuming the user will not input certain characters or will not expect the correct output for a foreign address –, the vocal database cannot be poorly designed.

The following sections describe the steps taken in the development of a high-quality speech corpus, with broad applications. The text selection mechanism is presented, as well as recording setup and speech annotation. The last section introduces a series of statistics based on the text used for the recordings.

3.3.1 Text Selection for the Recordings

The initial purpose of the speech corpus was to be as flexible and as extended as possible. The idea was to be able to use this corpus in a multitude of scenarios, from automatic speech recognition, to diphone and unit selection concatenation synthesis and of course for

HMM-based parametric system. One of the major requests of a concatenative diphone-based systems is the diphone coverage [4]. This requirement was achieved by recording a set of utterances which comprise multiple occurrences of the same diphone, and if possible all of the diphones of the language. The utterance selection for Romanian was performed using CDF. A list of the Romanian diphones with at least 10 occurences in the words of the DEX database was developed. The number of diphones used is 731. Each phoneme had to appear in 3 contexts (first, middle and end of word), and each context 3 times. CDF uses a greedy algorithm to select the best utterances, i.e. maximum number of required phonemes per utterance. Unfortunately, the algorithm failed to achieve these conditions because of the small text corpus. However, a set of 1000 utterances were selected[5]. A short excerpt of them is presented in Appendix F and combines two subsets, ***diph1*** and ***diph2*** with **500** utterances respectively.

A great advantage of the HTS system is the possibility to create a fairly natural synthetic speech with a limited amount of speech with almost no preprocessing or selection methods involved. In order to test this hypothesis, a set of **1500** random utterances selected from the newspaper text corpus was also recorded. This corpus is split into 3 subsets ***rnd1***, ***rnd2*** and ***rnd3***, each containing on average **500** utterances. Some of the random utterances are listed in Appendix G.

Initially, speech corpora for speech processing consisted of narrative texts, because of their availability and use of language. To comply with this specification two short fairytales ***"Povestea lui Stan Păţitul"*** and ***"Ivan Turbincă"*** by Ion Creangă, were also selected to be recorded. The choice was also made for the variation of prosodic patterns in the reading of such texts. Each of the two fairytales was split into utterances and read individually. They amount to **407** sentences for *Povestea lui Stan Păţitul* and **297** for *Ivan Turbincă*. A sample utterance segmentation is presented in Appendix H.

The sentences presented above, random, diphone coverage and the two fairytales, represent the training set for the speech synthesis system. For testing purposes, three additional sets of utterances have been developed. These include **210** random newspaper sentences, **110** fairytale sentences and **216** semantically unpredictable sentences. The fairytale sentences were selected at random from the narrative texts freely available

[4]The speech corpus must comprise all possible combinations of phonemes in a language
[5]An analysis of their diphone coverage is presented in section 3.3.6

at `http://ro.wikisource.org/wiki`. The semantically unpredictable sentences development is presented in the next section, as they represent an important part for the evaluation of a text-to-speech system.

3.3.2 Semantically Unpredictable Sentences

The Semantically Unpredictable Sentences (SUS) have been introduced as a compulsory part in the intelligibility evaluation of the TTS systems [Benoit et al., 1996]. The idea is that the human listener should not make educated guesses for a heard word based on the context, or the semantics of the phrase. For example if we would say *The grass is green*, even if the listener would not fully understand the word *green*, he would make a probabilistic assumption that the last word is *green*.

Given the hypothesis presented above, the SUS is a sentence which is *grammatically correct, but has no semantic meaning*. For example, *The rock listens to the garage.*

In order to create such a set of sentences, there are a few guidelines presented in [Benoit et al., 1996]. The first step is to determine a few semantic patterns for the future sentences. The patterns selected for Romanian are shown in Table 3.3 and the categories represent:

- Sbst - noun
- SbstM - masculine noun
- SbstF - feminine noun
- VbIntranz - intransitive verb
- VbTranz - transitive verb
- Prep - preposition
- Conj - conjunction
- WhWd - interrogative adverb, *cum, când, unde, cât*

For each of the semantic categories, a list of frequent words is then selected. The Romanian sets add up to 620 words. In each category the number of words is computed according to the equation below:

$$\{\textit{the number of words in each category}\} = \{\textit{the occurrence number of its category in the semantic patterns}\}\ X\ \{\textit{the number of sentences for each of the semantic patterns}\}$$

Table 3.3: Semantic patterns for the Romanian SUS. The last column represents the number of syllables already present in the sentence.

Word1	**Word2**	**Word3**	**Word4**	**Word5**	**Ending mark**	**No. of syllables**
{Sbst}	{VbIntranz}	{Prep}	{SbstM}	{AdjM}	.	0
{Sbst}	{VbIntranz}	{Prep}	{SbstF}	{AdjF}	.	0
{SbstM}	{AdjM}	{VbTranz}	{Sbst}	-	.	0
{SbstF}	{AdjF}	{VbTranz}	{Sbst}	-	.	0
{VbTranz}	{Sbst}	{Conj}	{Sbst}	-	.	0
{WhWd}	{VbTranz}	{Sbst}	{SbstM}	{AdjM}	?	0
{WhWd}	{VbTranz}	{Sbst}	{SbstM}	{AdjM}	?	0
{Sbst}	{VbTranz}	{Sbst}	care	{VbTranz}	.	2

The sentences are selected so as to minimise the length of the sentence. This is important because the memory of the listener should not be tested as well. An average number of 5 words/sentence has been shown to be sufficient in the context of the listening test. The length of the sentence is also given by the number of syllables, thus the necessity to specify the number of syllables already present in the sentence for the preselected words[6]. In the development of the sentences, also the words are not repeated, thus the relation presented above.

The complete list of the Romanian semantically unpredictable sentences can be found in Appendix I.

3.3.3 High Sampling Frequency Recordings

After the text selection for the recordings, the next step was of course recording it. The recordings were performed in an hemianechoic[7] chamber at the University of Edinburgh, Center for Speech Technology Research. Since the effect of microphone characteristics on HTS voices is still unknown, three high quality studio microphones were used: a Neumann u89i (large diaphragm condenser), a Sennheiser MKH 800 (small diaphragm condenser with very wide bandwidth) and a DPA 4035 (headset-mounted condenser). Fig. 3.1 shows the studio setup. All recordings were made at **96 kHz** sampling frequency and **24 bits per sample**, then down sampled to 48 kHz sampling frequency. This is a so-called

[6]Such as *care* in pattern 8.

[7]anechoic walls and ceiling, floor partially anechoic

Figure 3.1: Studio setup for recordings. Left microphone is a Sennheiser MKH 800 and the right one is a Neumann u89i. The headset has a DPA 4035 microphone mounted on it.

over-sampling method for noise reduction. The oversampling by a factor of 4 relative to the Nyquist rate (24 kHz) and down sample to 48 kHz, results in the improvement of the signal-to-noise-ratio by a factor of 4. For recording, down sampling and bit rate conversion, ProTools HD hardware and software was used.

The speaker is a native Romanian female. 8 sessions were conducted over the course of a month, recording around 500 utterances in each session. At the start of each session, the speaker listened to a previously recorded sample, in order to attain a similar voice quality and intonation. The prosody used for the diphone coverage and random sets was as flat as possible, while the fairytales were read using a more dynamic, narrative-like intonation style.

3.3.4 Speech Data Segmentation and Annotation

After the recordings were performed, the utterance level segmentation was conducted. All of the approximately 4 hours of recordings, both the training and the testing sets, were manually annotated at utterance level and segmented using Wavesurfer[8]. Some of the sentences were left aside due to clipping or incorrect pronunciations.

[8] `http://www.speech.kth.se/wavesurfer/`

Table 3.4: Phonetic coverage of each subset of the training corpus.

Subset	Sentences	Size [min]	Diphones	Diphones/ sentence	Quinph.	Quinph./ sentence
Random	1493	104	662	0.44	41285	27.5
Diphone	983	53	706	0.71	26385	26.3
Fairytale	704	67	646	0.65	29484	29.4

The resulted training speech corpus has **983** diphone coverage sentences with a total length of **53** minutes, **1493** random sentences of **104** minutes and the fairy tales have been segmented into **704** utterances amounting to **67** minutes.

The test corpus is comparably small, with a total duration of 28 minutes. It comprises **210** random newspaper utterance with a duration of 13 minutes, **110** randomly selected fairytale utterances – 8 minutes and **216** SUS amounting to 7 minutes.

Table 3.4 shows the total number of different diphones and quinphones in the training subsets. Diphones are the typical unit used for concatenative systems and quinphones are the base unit for HMM-based speech synthesis systems. A larger number of types implies

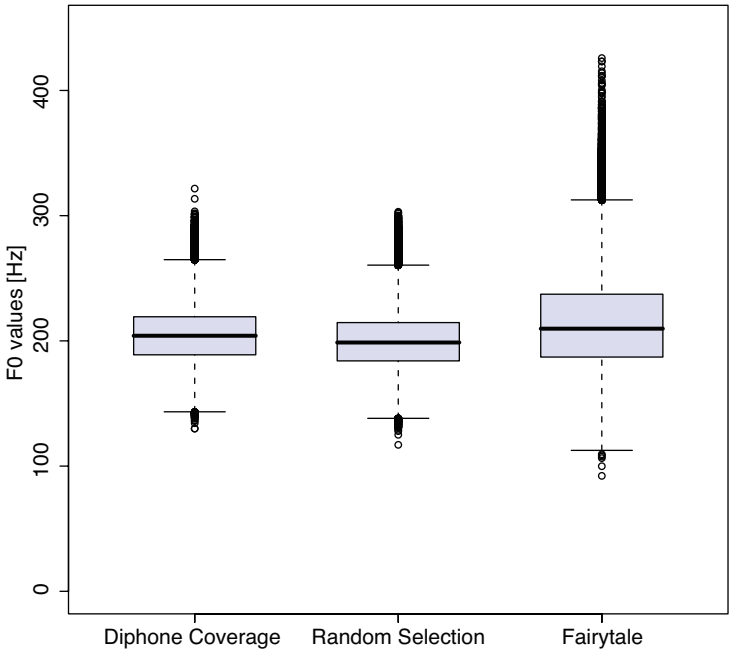


Figure 3.2: F_0 distributions in each training subset.

that the phonetic coverage is better. From the diphones/sentence column in the table we can see that the subset designed for diphone coverage has better coverage in terms of the number of different diphone types but – looking at the quinphones/sentence column – its coverage of quinphones is slightly worse than random selection. This indicates that the appropriate text design or sentence selection policy for HMM-based speech synthesis should be different from that for unit selection.

All recorded sentences were manually end-pointed and have been checked for consistency against the orthographic form. The newspaper sentences were read out using a relatively flat intonation pattern, while the fairy tales had a more narrative rhythm and prosody. Figure 3.2 shows the box-plots of F_0 values extracted from all the sentences of each of the training subsets, in which the mean is represented by a solid bar across a box showing the quartiles, whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. From this figure we can see that the subset including fairytales has a wider F_0 variation than the other subsets.

3.3.5 The Romanian Speech Synthesis (RSS) Corpus

The text and speech resources presented so far make up most of the structure of a freely available speech corpus entitled Romanian Speech Synthesis (RSS) corpus. Its structure is presented in Fig. 3.3. The corpus can be downloaded from `http://www.romaniantts.com/new/rssdb/rssdb.html` and includes the recordings for the training and testing sets, their phonetic transcription, the corresponding HTS labels (discussed in Chapter 4), accent positioning and samples of synthesised audio files using the TTS system developed within this thesis. The training HTS labels include time alignment, while the testing ones do not.

The aim of the corpus is for researchers with an interest in Romanian language to find a starting point resource for their applications. The corpus was tested within the built parametric synthesis system. A partial test within a concatenative synthesis system was also carried out (see Section 4.5).

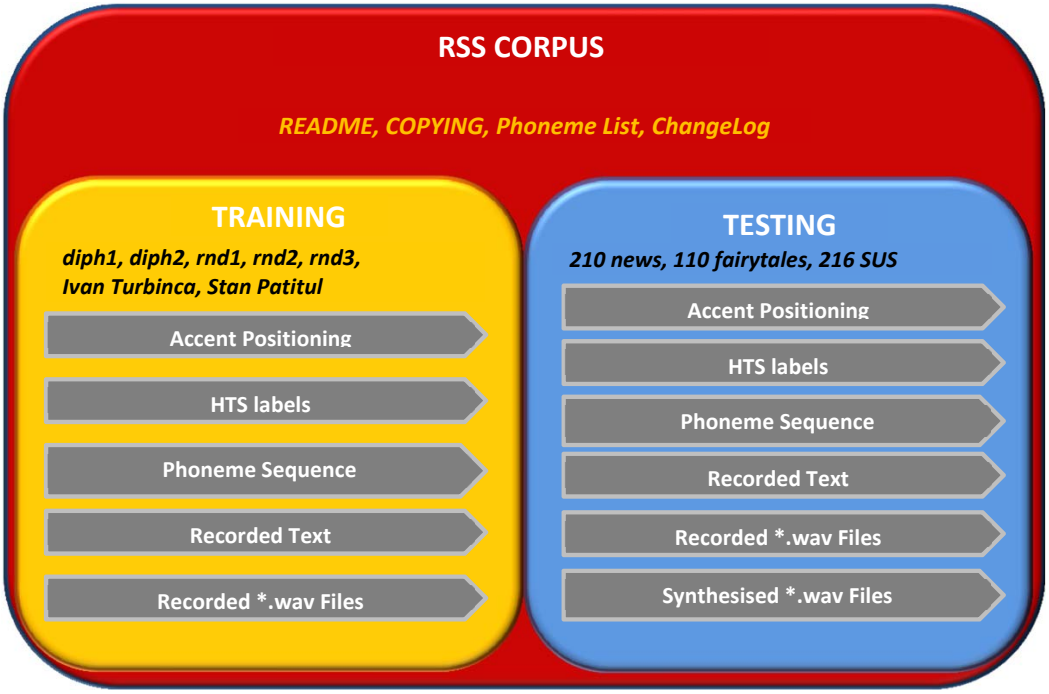


Figure 3.3: The structure of the Romanian Speech Synthesis (RSS) corpus.

3.3.6 Statistics of the Recorded Text in the RSS Corpus

To offer an insight to the textual elements of the RSS corpus, some statistics were employed using the **random** and **diphone** coverage training sets [Stan and Giurgiu, 2010]. These include: most frequent syllables, most frequent diphones, phoneme frequency and in the context of HTS, most frequent quinphones.

Most frequent syllables

Using the HTS labels generated for the random sentences and diphone coverage sets, the most frequent syllables were determined. Although there have not been any studies concerning the influence of a correct syllabification within the HTS labels, this information is used while building the clustering decision trees. Table 3.5 presents the top 40 most frequent syllables with their relative frequencies in the selected set.The accented characteristic is also presented, as in most of the European languages, the accent positioning can change the meaning of the word. There are a total of 2920 different syllables in the RSS text corpus and they add up to about 48,000 syllables. The statistic results are in

Table 3.5: The top 40 most frequent **syllables** and their relative frequencies in the selected speech corpus. The Accent column marks the accent of the syllable (0 - not accented, 1 - accented)

Syllable	**Accent**	**Frequency** [%]	**Syllable**	**Accent**	**Frequency** [%])
a	0	3.02	o	1	0.66
te	0	2.36	ta	1	0.61
de	1	2.13	ni	0	0.57
a	1	1.69	li	0	0.56
re	0	1.55	ţi	0	0.56
le	0	1.32	din	1	0.55
e	0	1.20	că	0	0.55
şi	1	1.19	pe	1	0.54
la	1	1.03	ce	1	0.53
ân	1	1.00	tru	0	0.50
ne	0	0.88	ti	0	0.50
nu	1	0.83	se	1	0.49
tă	0	0.78	mai	1	0.48
ca	1	0.76	ân	0	0.48
ri	0	0.75	me	0	0.47
de	0	0.72	au	1	0.46
ce	0	0.70	e	1	0.45
u	0	0.67	un	1	0.43
să	1	0.67	ma	0	0.43
cu	1	0.66	ră	0	0.43

correspondence with the ones obtained by [Buza, 2010] for an extended text corpus.

Phoneme frequencies

Given the importance that the phonemes have in the HTS parametric synthesiser, the relative frequency of the phonemes within the two sets was computed and is presented in Table 3.6.

It can be observed that the least frequent phonemes are the fricative *zh* and *dz* and the affricate *h*. These have been determined to cause some unnaturalness in the synthetic speech as well. So that, the RSS corpus should be enhanced with more samples of these 3 phonemes.

Table 3.6: Phoneme frequencies within the selected speech corpus.

Phoneme	**Frequency [%]**	**Phoneme**	**Frequency [%]**
e	10.64	ch	1.58
a	10.33	a@	1.49
i	7.09	v	1.38
r	6.78	sh	1.28
t	6.67	f	1.26
n	6.35	ij	1.14
u	5.58	ts	1.08
l	4.67	b	1.02
s	4.12	z	0.92
o	4.05	e@	0.86
k	3.74	w	0.73
m	3.39	g	0.69
p	3.18	o@	0.47
@	3.13	zh	0.31
d	3.10	dz	0.28
j	2.41	h	0.13

Most frequent diphones

Diphones have been one of the initial building blocks of concatenative synthesis, prior to unit selection. Their importance is still acknowledged in the current methods due to the effect of a phoneme over the following one. A proper diphone coverage in a speech corpus can determine improvements in the output quality. Table 3.7 presents the top 40 most frequent diphones in the selected speech corpus and their relative frequencies. The Romanian diphone inventory includes 731 diphones based on their occurrence in at least 10 words in the Romanian Explicative Dictionary (DEX) [DEX online-webpage, 2011]. The total number of diphones in the speech corpus is around 120,000.

Most frequent quinphones

In HMM-based speech synthesis the classification and regression trees are built on a series of features presented in Appendix D, of which one of the most important and

Table 3.7: The top 40 most frequent **diphones** and their relative frequencies in the selected speech corpus.

Diphone	**Frequency** [%]	**Diphone**	**Frequency** [%])
r-e	1.47	e-l	0.80
d-e	1.32	e@-a	0.80
t-e	1.28	t-a	0.79
a-r	1.28	a-n	0.77
a-t	1.17	a-l	0.76
i-n	1.16	k-u	0.75
a@-n	1.10	r-a	0.73
s-t	1.06	e-k	0.72
u-l	1.02	a-m	0.69
e-r	1.02	m-a	0.68
n-t	1.00	p-e	0.67
e-s	1.00	k-a	0.66
u-n	0.96	p-r	0.65
r-i	0.94	n-u	0.64
e-n	0.94	i-t	0.61
o-r	0.89	sh-i	0.57
t-r	0.86	n-i	0.57
l-e	0.83	e-d	0.56
l-a	0.82	u-r	0.55
ch-e	0.82	i-a	0.54

predominant in the question definition is the *phoneme context*[9]. This information is also known as a quinphone[10]. It is relevant while building a speech corpus for HTS to determine the best quinphone coverage. This is an impossible task, given that even in Romanian considering 32 phonemes plus silence and pause the possible quinphones add up to 270,000 possibilities.

In the selected subset there are around 57,000 different quinphones with 110,000 occurrences, which means that there is an approximate coverage of 25% with an average occurrence of 2. Table 3.8 presents the most frequent 40 quinphones and their relative frequencies within the set.

[9]The phoneme identity before the previous phoneme, the previous phoneme identity, the current phoneme identity, the next phoneme identity and the phoneme identity after the next phoneme

[10]There are 5 phonemes which determine the phoneme context.

Table 3.8: The top 40 most frequent **quinphones** and their relative frequencies within the selected speech corpus.

Quinphone	**Frequency** [%]	**Quinphone**	**Frequency** [%]
j-e-s-t-e	0.187	i-n-t-e-r	0.038
e-n-t-r-u	0.182	ts-j-o-n-a	0.038
p-e-n-t-r	0.177	a-m-e-n-t	0.038
a-ch-e-s-t	0.109	r-o-m-a-n	0.036
a-f-o-s-t	0.093	r-e-zh-e-ch	0.036
o@-a-r-t-e	0.073	s-p-r-e-z	0.036
f-o@-a-r-t	0.071	i-n-t-r-e	0.036
p-r-e-z-e	0.066	n-t-r-u-a	0.036
u-r-i-l-e	0.056	a-ch-e@-a-s	0.035
e-k-a-r-e	0.055	ch-e-a@-s-t	0.035
o@-a-m-e-n	0.048	ch-i-n-ch-ij	0.035
a-w-f-o-s	0.047	m-a-j-m-u	0.035
w-f-o-s-t	0.047	z-e-ch-ij-sh	0.035
r-i-l-o-r	0.045	e-ch-ij-sh-i	0.035
e-z-e-ch-e	0.044	f-@-k-u-t	0.035
t-u-l-u-j	0.042	s-p-e-k-t	0.033
t-a-t-e-a	0.041	a-j-m-u-l	0.033
t-r-e-b-u	0.041	t-o-r-u-l	0.033
a@-n-ch-e-p	0.039	e-p-e-n-t	0.033
s-p-u-n-e	0.038	p-a-t-r-u	0.032

3.4 Summary

This chapter introduced the development of several Romanian text and speech resources. These resources are an essential prerequisite for developing a Romanian text-to-speech system. In the first section, a short overview of the necessities for a correct resource acquisition and development is presented. These include the quality of the resulted system, the correct phonetic transcription or broad use of the resources. The entire chapter is built around two main units: *the text resources* and *the speech resources*.

Text resources were developed according to the requirements of a simple TTS front-end, and include:

- a collection of 4506 newspaper articles with over 1,700,000 words trawled from the online newspaper "Adevărul"
- a simplified phonetic inventory for Romanian which comprises 32 phonemes
- a set of basic rules for phonetic transcription written in Festival
- a source for correct accent positioning identified as the SQL database of the Online Romanian Explicative Dictionary (DEX)
- a preliminary evaluation of the Maximal Onset Principle of syllabification applied to Romanian
- part-of-speech tagging of the entire text corpus using [Sabou et al., 2008]
- a 65,000 word lexicon with correct phonetic transcription and accent positioning

The nature of the text resources is rather general than particular. One aspect left aside within this approach is the Romanian text normalisation, which is more of a Natural Language Processing problem than a text-to-speech one.

A very important addition to the available resources is the speech corpus. Given the prior evaluation of the freely available Romanian speech resources, the need for an extended, broad application speech corpus was identified. This lead to the acquisition of approximately 4 hours of high-quality recordings which include text from both newspapers and narrative writings. They include:

- 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. 1/2 hour

  - 210 random newspaper utterances
  - 110 random fairytale utterances
  - 216 semantically unpredictable sentences

The recordings were done in a hemianechoic room with 3 simultaneous microphones, at 96kHz sampling frequency and 24 bits per sample. To achieve an even lower level of noise, the speech was down sampled at 48kHz, 24 bits per sample.

Another important development is the set of 216 semantically unpredictable sentences, essential in the evaluation of a TTS system and unavailable for Romanian. The development process is presented in section 3.3.2 and the complete list in Appendix I.

A selection of the resources was included in the Romanian Speech Synthesis database, freely available online. For a correct evaluation of the RSS database, statistics of the recorded text, such as phoneme or diphone relative frequencies, are briefly presented in section 3.3.6.

All of the presented resources are freely available, some on request, such as the lexicon or the text corpus and some online, at `www.romaniantts.com` within the RSS database. They provide a starting point for the comparative evaluation of TTS systems for example, or the development of new and enhanced resources or speech processing systems, be them automated speech recognition or text-to-speech synthesis.

Chapter 4

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

4.1 Introduction

HMM-based statistical parametric speech synthesis [Zen et al., 2009] has been widely studied and has now become a mainstream method for text-to-speech systems. The HMM-based speech synthesis system (HTS) [Zen et al., 2007a] is the principal framework that enables application of this method to new languages. It has the ability to generate natural-sounding synthetic speech and, in recent years, some HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems [Karaiskos et al., 2008] in terms of naturalness and intelligibility. However, relatively poor perceived "speaker similarity" remains one of the most common shortcomings of such systems [Yamagishi et al., 2008b].

One possible reason that HMM-based synthetic speech sounds less like the original speaker compared to a concatenative system built from the same data, may be the use of a vocoder, which can cause buzziness or other processing artefacts. Another reason may be that the statistical modelling itself can lead to a muffled sound, presumably due to the

process of averaging many short-term spectra, which removes important detail.

In addition to these intrinsic reasons, there are also extrinsic problems: some basic configuration choices in HMM synthesis have been simply taken from different fields such as speech coding, automatic speech recognition and unit selection synthesis. For instance, 16 kHz is generally regarded as a sufficiently high waveform sampling frequency for speech recognition and synthesis because speech at this sampling frequency is intelligible to human listeners. However speech waveforms sampled at 16 kHz still sound slightly muffled when compared to higher sampling frequencies. HMM synthesis has already demonstrated levels of intelligibility indistinguishable from natural speech [Karaiskos et al., 2008], but high-quality TTS also needs to achieve naturalness and speaker similarity.

Another practical, but equally important, factor is footprint. In unit selection, higher sampling frequencies may lead to a larger footprint. However, the use of higher sampling frequencies does not in itself change the footprint of a HMM-based speech synthesis system. The use of higher sampling frequencies increases computational costs for both methods.

These apparently basic issues are revisited within this chapter in order to determine whether current configurations are satisfactory, especially with regard to speaker similarity. As the sampling frequency increases, the differences between different auditory frequency scales such as the Mel and Bark scales [Zwicker and Scharf, 1965] implemented using a first-order all-pass function become greater. The experiments also included a variety of different auditory scales.

The RSS corpus is also evaluated and the best set of training data for a qualitative synthesis is determined. The training set is a key feature in providing the statistical models with sufficient data. Given the speech corpus presented in the previous chapter, several combinations of its subsets are used to build HTS voices.

Section 4.5 reports the results of a Blizzard-style listening tests [Karaiskos et al., 2008] used to evaluate HMM-based speech synthesis using higher sampling frequencies as well as standard unit selection voices built. The results suggest that a higher sampling frequency can have a substantial effect on HMM-based speech synthesis.

This chapter is organised as follows. Section 4.2 presents the theoretical aspects of the parametric HMM-based speech synthesiser. Section 4.3 describes the prerequisites HTS

system, from text, to decision questions and speech data, while section 4.4 defines the configuration for the HTS parameters. In section 4.5 the evaluation of the resulted system using a Blizzard-style listening test and the afferent results are presented. As an additional and continuous evaluation, the system is available for an interactive demonstration online, described in section 4.5.2. A side experiment using the adaptation capabilities of HTS is presented in Section 4.5.3.

4.2 HMM-based Speech Synthesis

4.2.1 The Hidden Markov Model

A hidden Markov model is a finite state machine which generates time discrete observations. In a Markov chain, each state corresponds to a deterministic observable event. Non-deterministic processes are the input of the hidden state models and the output is any of model's state. So that, an observation is a state dependent probabilistic function. It therefore exists a hidden stochastic process which cannot be observed. The hidden process can only be associated with another observable process, producing a series of observable characteristics. At each time sample, HMM modifies its states according to a transition probability and generates the observation o according to the probability distribution of the current state. A continuous HMM is described according to [Yamagishi, 2006] by:

- o - an output observation data. The observation data corresponds to the physical output of the system being modelled
- $\omega = 1, 2, ...N$ - a set of states representing the state space. Here s_t is denoted as the state at time t
- $A = a_{ij}$ - a transition probability matrix, where a_{ij} is the probability of taking a transition from state i to state j, i.e. $a_{ij} = P(s_t = j|s_{t-1} = i)$
- $B = b_i(o)$ - an output probability distribution. The output probability distribution $b_i(o)$ of the observational data o of state i is modelled by a mixture of multivariate Gaussian distributions according to $b_i(o) = \sum_{m=1}^{M} w_{im}\mathcal{N}(o; \mu_{im}, \Sigma_{im})$, where M is the number of mixture components of the distribution, and w_{im}, μ_{im} and Σ_{im} are a weight, a L-dimensional mean vector, and a $L \times L$ covariance matrix of mixture component m of state i, respectively. A Gaussian distribution $\mathcal{N}(o; \mu_{im}, \Sigma_{im})$ of each

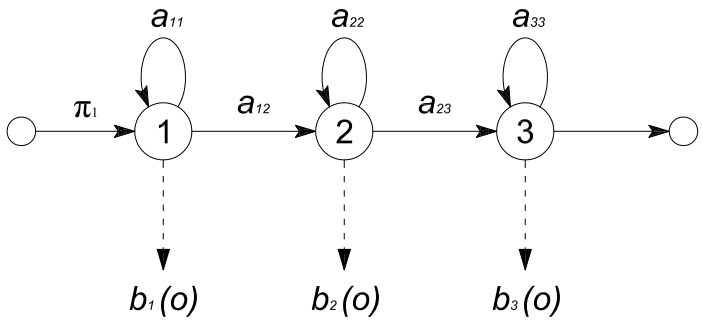


Figure 4.1: Example of a left to right HMM structure

component is defined by $\mathcal{N}(o;\mu_{im},\Sigma_{im}) = \frac{1}{(2\pi)^L|\Sigma_{im}|} exp\left(-\frac{1}{2}(o-\mu_{im})^\top \Sigma_{im}^{-1}(o-\mu_{im})\right)$, where L is the dimensionality of the observation data o.

- $\pi = \pi_i$ an initial state distribution where $\pi_i = P(s_o = i), 1 \leqslant i \leqslant N$

The following properties must be satisfied:

$$a_{ij} \geqslant 0, w_{im} \geqslant 0, \pi_i \geqslant 0, \forall i,j,m$$
$$\sum_{j=1}^{N} a_{ij} = 1,\ i = 1...N$$
$$\sum_{m=1}^{M} w_{im} = 1,\ i = 1...N$$
$$\sum_{i=1}^{N} \pi_i = 1,\ i = 1...N$$
$$\int_o b_i(\mathbf{o}) d\mathbf{o} = 1$$

To sum up, a complete specification of an HMM includes two constant-size parameters, N and M the total number of states and the number of mixture components, w_{im} the Gaussians weights, the observational data o, and three sets (matrices) of probability measures A, B, π in the following notation:

$$\phi = (A, B, \pi)$$

to indicate the whole parameter set of an HMM.

In speech processing, the most common HMM structure used is the left-right one (Fig. 4.1). In this structure, the state index is incremented or remains constant. This type of model approximated correctly the speech signal, whose characteristics modify over time.

4.2.2 Speech Signal Parameter Modelling

As a parametric synthesis method, HMM-based speech synthesis needs a set of features extracted from the speech in order to estimate its inner models. In the HMM-based speech synthesis, the speech parameters of a speech unit such as the spectrum, fundamental frequency (F0), and phoneme duration are statistically modelled and generated by using HMMs based on maximum likelihood criterion [Yamagishi, 2006]. The speech is analysed at frame level and develops context-based models for each phoneme. The spectrum features are represented by the Mel-Frequency Cepstral Coefficients, similar to the method used in automatic speech recognition. The following subsections describe the extraction of the feature vector and the building of the HMMs.

Mel-cepstrum analysis

In speech analysis, the most common model for speech production is the source-filter model. Within the mel-cepstrum analysis, the transfer function of the vocal tract, $H(z)$ is modelled by the mel-cepstrum coefficients (MFCC). This representation is obtained by applying the discrete Fourier transform (DFT) over a speech frame. The Fourier spectrum is then filtered through a Mel-scale frequency filterbank. From each sub band, the log of the power is computed and the discrete cosine transform (DCT) is applied to the result. The MFCCs are the amplitudes of the resulting spectrum(Fig. 4.2). The correspondence between mel-scale and normal frequency scale is:

$$\mathrm{m} = 2595 log_{10} \left(1 + \frac{f}{700}\right)$$

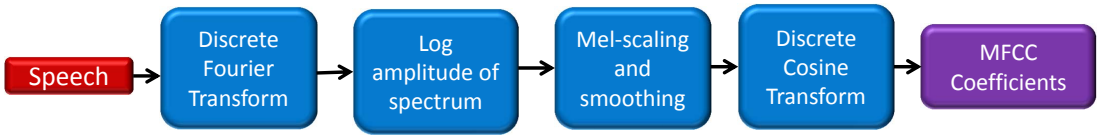


Figure 4.2: MFCC coefficients computation

The advantage behind this type of representation is the fact that these coefficients remain independent and allow for a probability distribution modelling by a diagonal covariance matrix. Along with the MFCC coefficients, the feature vector also includes the delta and delta-delta coefficients of the MFCC.

Fundamental Frequency Modelling

MFCC models the spectrum of the speech, while an important characteristic of speech is the pitch or fundamental frequency. Because of the lack of pitch values in the unvoiced segments, F0 cannot be modelled using conventional discrete or continuous HMMs. Thus, a new type of HMMs are defined, the Multi-space Probability Distribution HMM (MSD-HMM) [Tokuda et al., 1999], ([Tokuda et al., 2002a]). In order to model the pitch using MSD-HMMs two spaces are defined: a one dimensional space with a probability density function for the voiced segments and a zero dimensional space containing a single point for the unvoiced segments. In this way, F0 can be modelled without making any heuristic assumptions of its values.

HMM state duration modelling

In standard HMM models, the transition probabilities determine the duration characteristics of the model. In phoneme synthesis, the duration must be explicitly specified, because of its major influence in the speaker characteristics and in the rhythm and prosody of speech. Another type of HMM models is so defined. They are called Hidden Semi-Markov Models (HSMM) and the transition probabilities are replaced by explicit Gaussian models for the duration.

4.2.3 Decision Tree Building for Context Clustering

In continuous speech, the parameter sequences of an acoustic unit varies according to the phonetic context. The correct modelling of these variations implies context dependent models, such as triphones or quinphones. In HMM-based speech synthesis systems, the context is defined by both the phonetic and the linguistic and prosodic context. A contextual clustering is achieved using binary decision trees. Each tree node defines a cluster based on a contextual factor. Each tree leaf contains an output probability distribution

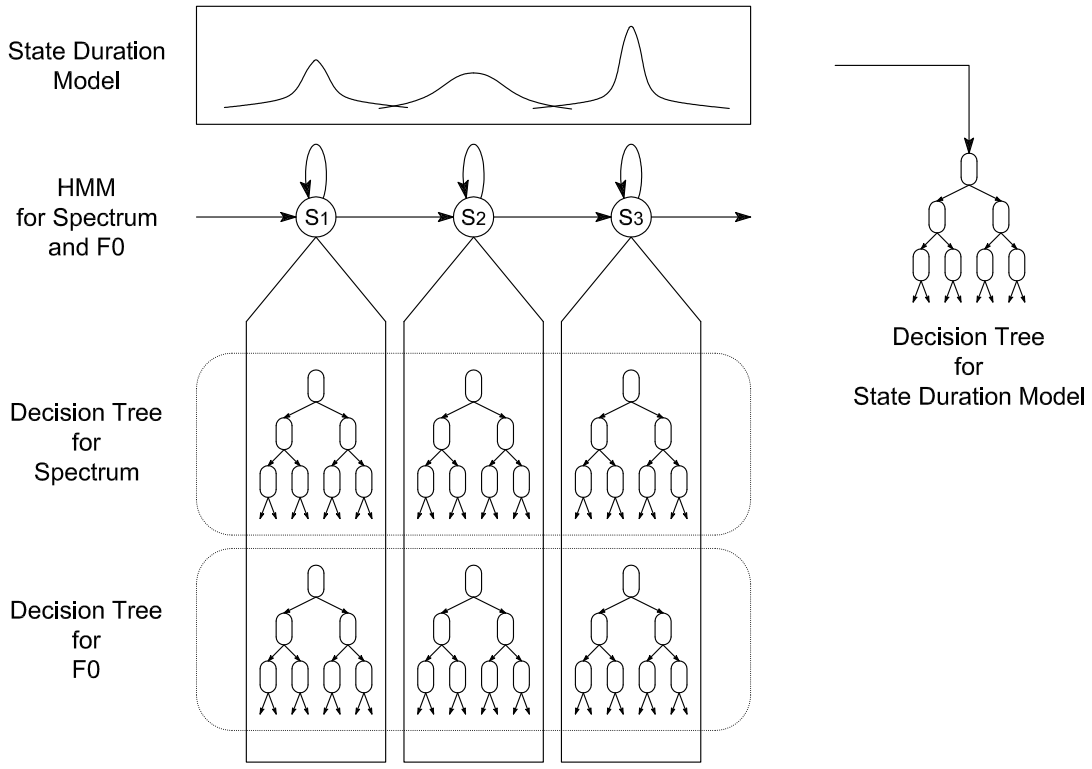


Figure 4.3: Decision tree context clustering in HMM-based speech synthesis (after [Tokuda et al., 2002b]).

of the state. The trees are built using the Minimum Description Length (MDL) principle are are used to cluster pitch, duration and spectrum. Fig. 4.3 shows the three different decision trees built for context clustering in HTS.

4.2.4 Speech Parameter Generation

The input labels of the HMM-based speech synthesiser offer information about the phoneme sequence, but not about the HMM states that should be used in synthesis. To determine the state sequence, the Maximum Likelihood (ML) algorithm is applied. The MFCC coefficients are synthesised using a Mel Log Spectrum Approximation (MLSA) filter [Imai et al., 1983].

4.2.5 The HMM-based Speech Synthesis System

The HMM-based Speech Synthesis System (H Triple S - HTS) is a collection of open source tools dedicated to the development of text-to-speech systems using Markov models [HTS webpage, 2010]. The content of these tools refer strictly to the modelling, training and speech generation without text processing. The system input is represented by the HTS labels presented in section 4.3.1.

HTS is built on the Hidden Markov Model Toolkit (HTK) [Young et al., 2001]. HTK was initially developed for automatic speech recognition. The training part of the HTS is a modified version of HTK. In Fig. 4.4 the block diagram of the HTS system is presented. Two main sections can be observed: training and synthesis. In the training section, the spectrum, pitch and duration HMM models are extracted. Decision tree clustering is applied to the MFCC coefficients, pitch values and duration. The HMMs are re-estimated using a Baum-Welch algorithm. The result of the training section are the decision tree clusters with their respective parameters in the leaf nodes.

The synthesis section uses HTS labels to generate phoneme level HMM state sequences. The MLSA filter is then applied to generate the synthetic speech from the state parameters.

HTS is very flexible and allows for the following parameter modification, both in the training and in the synthesis sections:

- training data set
- sampling frequency
- non-linear transformation of the frequency scale
- analysis/synthesis frame length
- cepstral order
- analysis method: STRAIGHT Mel-cepstrum, STRAIGHT Mel-generalised cepstrum, STRAIGHT Mel-LSP, STRAIGHT Mel-generalised LSP
- pitch estimation method: IFAS, Fixed-Point analysis, ESPS, voting between previous methods
- number of HMM states
- the root node decision tree question

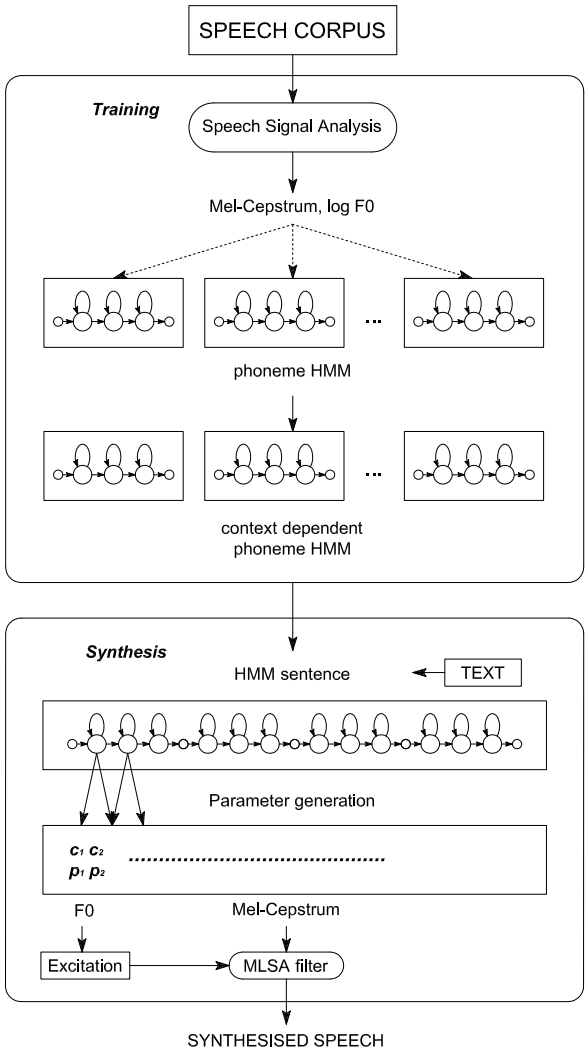


Figure 4.4: Basic HTS structure (after [Yamagishi, 2006]).

An important development of the HTS system is the speaker conversion described by [Stylianou et al., 1998, Ohtani et al., 2006]. Starting from the speaker dependent or independent trained decision trees, using Maximum Likelihood Linear Regression (MLLR), the models are adapted to a new training data. The results of the adaptation are not considered high-quality, but the amount of necessary training data is largely reduced [Yamagishi et al., 2009]. Speaker adaptation represents one of the major advantages of parametric synthesis over the concatenative system. Studies show that an average of 5 minutes recordings can capture the gross features of a new speaker if the starting models are speaker independent.

4.3 Data Preprocessing

4.3.1 Prerequisites to an HTS Compliant Text Annotation

As the scope of this thesis relies mostly on the speech synthesis, and less on the text processing, a very basic, simple Romanian text processor was developed using the technology provided by Cereproc [Aylett and Pidcock, 2007]. The main focus was the ability to create HTS format labels from raw text. Text normalisation was not taken into account, and letter-to-sound rules are simplified. The text resources described in Chapter 3 represent the basis of the text processor. CDF is a commercial tool and intrinsic aspects of the implementation are not public. At the point of the implementation, Cereproc offered mainly concatenative systems, so that the output of the front-end had to be converted into HTS specific labels.

The HMM-based Speech Synthesis System requires a very elaborate text annotation in order to correctly classify the phoneme features based on the context they appear in. The complete list of features is presented in Annex D and are referred to as HTS labels. It can be noted that most of the features require an extensive text processing and can derive problems when not involving a correct text processor. For the system training, the HTS labels also include the temporal markers for the beginning and end of the phoneme, while in the synthesis part, the duration is given by the decision tree model.

From the full set of features, a few have been left aside or had a reduced form, due to the lack of knowledge and resources available. These are:

- ToBI labels - require manual annotation of the F0 contour for the training stage. The speech corpus developed and used did not benefit from any manual annotation therefore, the default value for the ToBI labels in the training set is L-L. In the synthesis part, although the text processor could derive heuristic ToBI labels, the decision trees of the HTS voice are not trained in such manner, thus resulting the same intonation no matter the ToBI labels.
- Part-of-Speech tags - have been reduced to the *feature*, *content* categories because of the lack of accuracy of the POS tagger used[1].

[1]The authors reported in an offline document an average accuracy of 70%

- Stress - in Romanian there is no difference between stress and accent, so that the stress marker equals the accent marker for the Romanian labels
- Name of the vowel - was considered the vowel within the syllable and not the diphthong or triphthong

Their influence in the result of the system has not yet been evaluated, but given the broad spectrum of features used, they are not considered as essential for the quality of the synthetic speech.

4.3.2 Decision Tree Questions for Romanian

The HMM models are clustered using the binary decision trees. The tree nodes are defined by the significant context of the current phoneme and influence the clustering in the training stage. The correct definition of the nodes' questions is therefore of great importance. The HTS labels are built using the full context of the phoneme, and the features are presented in Appendix D. All of the features are used in the final decision tree, but their influence is weighted according to the decision tree building algorithm.

Some of the context features are language dependent, such as the phonetic context, or the name of the vowel. But the rest of them are language independent and represent for example the number of syllables before the current syllable or the number of words in the phrase. For the phonetic set of features the questions have to be rewritten for the Romanian set of phonemes. Below is a short sample of the questions defined for the Romanian HTS system and are built using the Unilex sample questions for English:

QS "LL-Trill"	{r~* }
QS "LL-Approximant"	{l ~* }
QS "LL-BilabialNasal"	{m~*}
QS "LL-DentalNasal"	{n~*}
QS "LL-Bilabial_Plosive"	{p~*, b~* }
QS "L-m"	m~*
QS "L-n"	n~*
QS "L-f"	f~*

QS "L-v"	v~*
QS "L-s"	s~*
QS "C-Nasal"	*-m+*,*-n+*
QS "C-Plosive"	*-p+*,*-t+*,*-k+*,*-b+*,*-d+*,*-g+*
QS "C-Voiced_Plosive"	*-b+*,*-d+*,*-g+*
QS "C-Unvoiced_Plosive"	*-p+*,*-t+*,*-k+*
QS "C-Affricates"	*-ts+*,*-ch+*,*-dz+*
QS "R-Front_Vowel"	*+i=*,*+e=*,*+e@=*,*+ij=*,*+j=*
QS "R-Front_close_vowel"	*+i=*,*+ij=*,*+j=*
QS "R-Front_mid_vowel"	*+e=*,*+e@=*
QS "R-Front_nearback_vowel"	*+iw=*,*+ew=*,*+we=*,*+jew=*
QS "R-Front_nearfront_vowel"	*+ij=*,*+ej=*,*+je=*,*+jej=*,*+jew=*
QS "RR-o@"	*=o@:*
QS "RR-u"	*=u:*
QS "RR-w"	*=w:*
QS "RR-@"	*=@:*
QS "RR-a@"	*=a@:*

It can be observed that the detail level in the definition of the questions is quite high. Apart from the vowel/consonant categories, each of the phonemes are described through their sound quality. For the phonetic context alone there are 712 questions defined, repeated for the right-right, right, current, left and left-left contexts. This means a number of 178 distinct questions for each position.

4.3.3 Prerequisites to an HTS Compliant Speech Corpus

As opposed to unit selection in concatenative speech synthesis, HMM-based parametric systems require a less extended speech corpus. A major issue relies in the acquisition of enough examples of phonemes in certain contexts. Having the quinphone as the basic unit of synthesis, the correct coverage of the quinphones is necessary although not essential. The correlation effects of the phoneme pronunciation may lead to some contexts being clustered in the same class, no matter how many samples of speech are available. Although, it is important to maintain a flat intonation throughout the corpus, thus limiting

the number of context clusters and ensuring enough training samples within them.

The only major requirement is that the speech corpus is labeled using the HTS format and that the training labels contain phoneme level time alignment. Having the text processor already available, the entire speech corpus was labelled. All the words found in the recorded sentences were checked in the lexicon for correct phonetic transcription. The time alignment was carried out in a bootstrapping manner using the result labels of the HTS training step. The initial time alignment was simply determined by dividing the total length of the utterance to the number of phonemes in it. The training labels have the format presented in Appendix D.

4.4 Building an HMM-based Speech Synthesis System using High Sampling Frequency

A recent HMM-based speech synthesis system described in [Zen et al., 2007b] was adopted. It uses a set of speaker-dependent context-dependent multi-stream left-to-right state-tied [Young et al., 1994, Shinoda and Watanabe, 2000] multi-space distribution (MSD) [Tokuda et al., 2002a] hidden semi-Markov models (HSMMs) [Zen et al., 2007c] that model three kinds of parameters, required to drive the STRAIGHT [Kawahara et al., 1999] mel-cepstral vocoder with mixed excitation [Kawahara et al., 2001]. Once the context-dependent labels from the language-dependent front-end outputs are defined, the framework of this system is basically language-independent and thus it can be directly used on the data.

The sampling frequency of the speech directly affects feature extraction and the vocoder and indirectly affects HMM training via the analysis order of spectral features. The following sections give an overview of how the sampling frequency affects the first-order all-pass filter used for mel-cepstral analysis and how the higher sampling frequencies in this analysis method can be utilised.

4.4.1 The first-order all-pass frequency-warping function

In mel-cepstral analysis [Tokuda et al., 1991], the vocal tract transfer function $H(z)$ is modelled by M-th order mel-cepstral coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\top$ as follows:

$$H(z) = \exp \boldsymbol{c}^\top \tilde{\boldsymbol{z}} = \exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m}, \tag{4.1}$$

where $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \ldots, \tilde{z}^{-M}]^\top$. $\tilde{z}^{-1}$ is defined by a first-order all-pass (bilinear) function

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1 \tag{4.2}$$

and the warped frequency scale $\beta(\omega)$ is given as its phase response:

$$\beta(\omega) = \tan^{-1} \frac{(1 - \alpha^2) \sin \omega}{(1 + \alpha^2) \cos \omega - 2\alpha}. \tag{4.3}$$

The phase response $\beta(\omega)$ gives a good approximation to an auditory frequency scale with an appropriate choice of α.

An example of frequency warping is shown in Fig. 4.5. where it can be seen that, when the sampling frequency is 16 kHz, the phase response $\beta(\omega)$ provides a good approximation to the mel scale for $\alpha = 0.42$. The choice of α depends on the sampling frequency used and the auditory scale desired. The next section describes how to determine this parameter for a variety of auditory scales.

4.4.2 The Bark and ERB scales using the first-order all-pass function

In HMM-based speech synthesis, the mel scale is widely used. For instance, Tokuda *et al.* provide appropriate α values for the mel scale for speech sampling frequencies from 8kHz to 22.05kHz [Tokuda et al., 1994a].

In addition to the mel scale, the Bark and equivalent rectangular bandwidth (ERB) scales [Patterson, 1982] are also well-known auditory scales. In [Smith III and Abel, 1999],

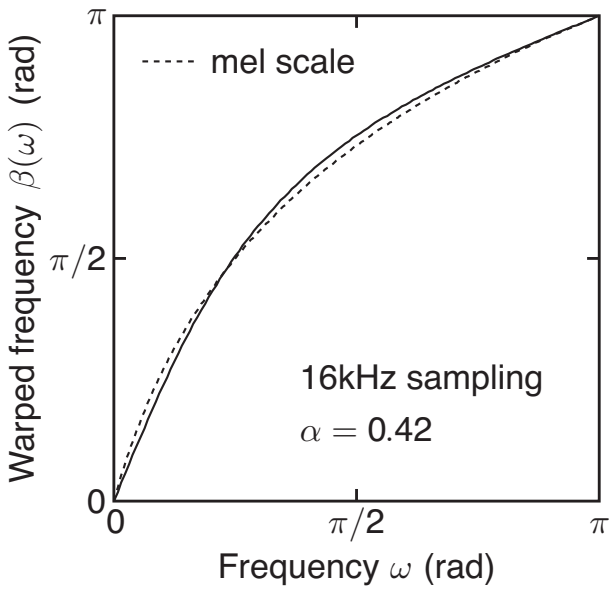


Figure 4.5: Frequency warping using the all-pass function. At a sampling frequency of 16 kHz, $\alpha = 0.42$ provides a good approximation to the mel scale.

Smith and Abel define the optimal α (in a least-squares sense) for each scale as follows:

$$\alpha_{\text{Bark}} = 0.8517\sqrt{\arctan(0.06583\, f_s)} - 0.1916 \tag{4.4}$$

$$\alpha_{\text{ERB}} = 0.5941\sqrt{\arctan(0.1418\, f_s)} + 0.03237 \tag{4.5}$$

where f_s is the waveform sampling frequency. However, note that the error between the true ERB scale and all-pass scale approximated by α_{ERB} is three times larger than the error for the Bark scale using α_{Bark} [Smith III and Abel, 1999]. Note also that as sampling rates become higher, the accuracy of approximation using the all-pass filter becomes worse for both scales.

4.4.3 HMM training

The feature vector for the MSD-HSMMs consists of three kinds of parameters: the mel-cepstrum, generalised $\log F_0$ [Yamagishi and King, 2010] and a set of band-limited aperiodicity measures [Ohtani et al., 2006], plus their velocity and acceleration features.

An overview of the training stages of the HSMMs is shown in Figure 4.6. First, monophone MSD-HSMMs are trained from the initial segmentation using the segmental K-

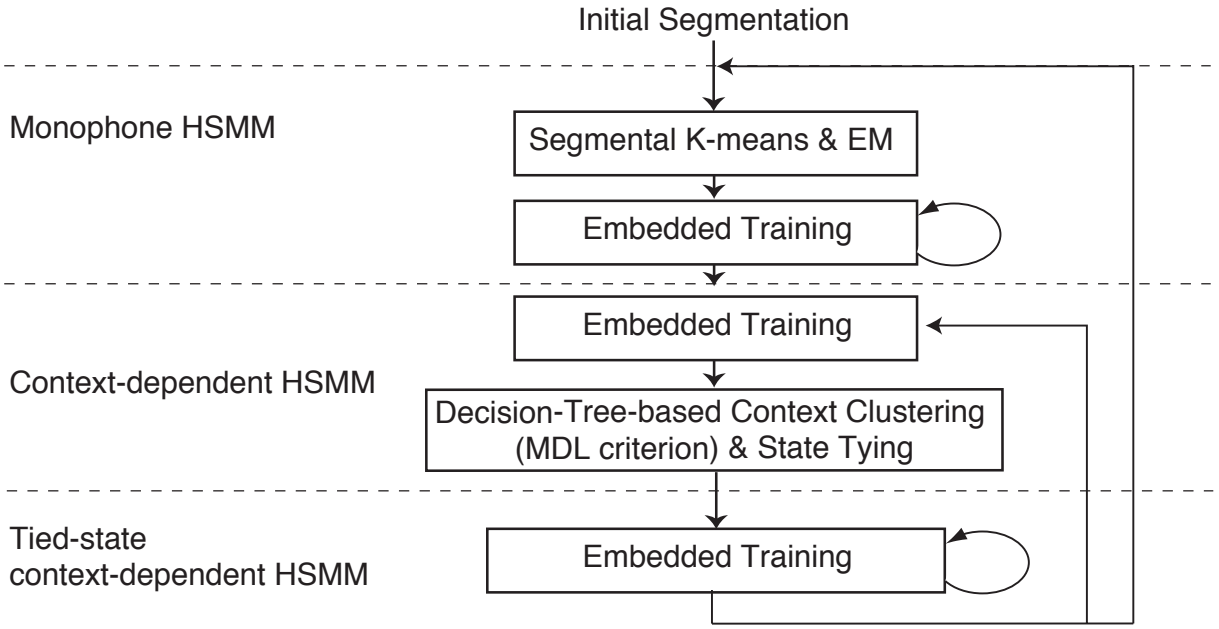


Figure 4.6: Overview of HMM training stages for HTS voice building.

means and EM algorithms [Dempster et al., 1977], converted to context-dependent MSD-HSMMs and re-estimated using embedded training. Then, decision-tree-based context clustering [Young et al., 1994, Shinoda and Watanabe, 2000] is applied to the HSMMs and the model parameters of the HSMMs are thus tied. The clustered HSMMs are re-estimated again using embedded training. The clustering processes are repeated until convergence of likelihood improvements (inner loop of Figure 4.6) and the whole process is further repeated using segmentation labels refined with the trained models in a bootstrap fashion (outer loop of Figure 4.6). In general, speech data sampled at higher rates requires a higher analysis order for mel-cepstral analysis. Therefore the process started by training models on lower sampling rate speech (16 kHz) with a low analysis order and gradually increased the analysis order and sampling rates via either re-segmentation of data or single-pass retraining of HMMs [Yamagishi and King, 2010].

4.4.4 Configurable parameters

In order to establish a benchmark system which will be useful for many future experiments, the various configurable parameters were carefully adjusted as follows:

1. From initial analysis-by-synthesis tests using five sentences followed by informal listening, the spectral analysis method and order are chosen. Specifically, the mel-

cepstrum and mel-generalised cepstrum (MGC) [Tokuda et al., 1994b] at orders of 50, 55, 60, 65 and 70, using Bark and ERB frequency warping scales[2] using speech data sampled at 48 kHz were compared. The parameter to control all-pole or cepstral analysis method was set to 3 [Tokuda et al., 1994b]. The results indicated the use of MGC with 60th order and the Bark scale. However, the differences between the Bark and ERB scales were found to be not as great differences due to the sampling frequency. [Yamagishi and King, 2010] also found that the auditory scale – including the Mel scale – was not a significant factor. Therefore the ERB scale and the Mel scale were omitted from the listening test reported later. The same process for speech data sampled at 32 kHz and chose MGC with 44th order with the Bark scale was repeated.

2. Preliminary HMM training was then carried out to determine training data partitions. A total of 20 systems resulted from combinations of the recorded data used in sets of 500, 1000, 1500, 2500 and 3500 sentences. From informal listening, the fairy tale sentences were found to alter the overall quality of the synthesised speech, since these sentences had a more dynamic prosody than the newspaper sentences (see Figure 3.2). Therefore the fairy tale set was excluded and a 2500 sentence set was used in subsequent experiments.

3. The data-driven generalised-logarithmic F_0 scale transform method proposed in [Yamagishi and King, 2010] was employed. The maximum likelihood estimator for the generalised logarithmic transform obtained from F0 values of all voiced frames included in the RSS database is **0.333**, calculated using the optimisation method described in [Yamagishi and King, 2010], .

4. The decision trees for speech from non-speech units (pauses and silences) were separated using as root tree question $C - sil$, rather than having a shared single tree.

In the experiments reported in this chapter, only speech recorded using the Sennheiser MKH 800 microphone was used. Investigation of the differences caused by the microphone type are left as future work.

[2]Strictly speaking, they should be called Bark-cepstrum and ERB-cepstrum. However, for simplicity they will all be called 'mel-cepstrum'.

Table 4.1: Mean scores for the speech synthesis listening test sections

	A	B	C	D	E	F	G	H	I
Similarity	4.9	2.6	2.5	2.7	3.1	3.0	3.1	3.4	3.3
Naturalness	4.8	2.2	2.5	2.4	3.3	3.0	3.4	3.4	3.4
Intelligibility (WER [%])	1.0	5.0	5.8	7.1	4.1	8.0	5.0	3.5	4.5

4.5 Evaluation

4.5.1 Experiment 1 – Listening Test

For the listening test, the framework from the Blizzard Challenge [Karaiskos et al., 2008] was used, and evaluated speaker similarity, naturalness and intelligibility.

A total of 54 Romanian native listeners were recruited, of which 20 completed the test in purpose-built, soundproof listening booths and the rest evaluated the systems on their personal computers and audio devices, mostly using headphones. They each evaluated a total of 108 sentences randomly chosen from the test set, 36 from each category (news, novel, SUS). The speaker similarity and naturalness sections contained 18 newspaper sentences and 18 novel sentences each. 36 SUSs were used to test intelligibility.

The duration of the listening test was about 45 minutes per listener. Listeners were able to pause the evaluation at any point and continue at a later time, but the majority opted for a single listening session. Most of the listeners had rarely listened to synthetic voices; they found the judgement of naturalness and speaker similarity to be the most challenging aspects of the test.

Nine individual systems were built for the evaluation. All used the same front-end text processing. They differ in the synthesis method used (HMM-based, unit selection), sampling frequency (16 kHz, 32 kHz, 48 kHz) and the amount of data used for the training of the voice. The analysis of the three microphones is an interesting topic but, in order to make the listening tests feasible, this factor had to be excluded. The systems are identified by letter:

A Original recordings, natural speech at 48 kHz

B Unit selection system at 16 kHz, using 3500 sentences

C Unit selection system at 32 kHz, using 3500 sentences

Table 4.2: Significance at 1% level for (a) **similarity** , (b) **naturalness** and (c) **WER**, based on Wilcoxon signed rank tests with alpha Bonferoni correction (1% level); '1' indicates a significant difference.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	1	1	1	1	1
C	1	0	-	0	1	1	1	1	1
D	1	0	0	-	1	0	0	1	1
E	1	1	1	1	-	0	0	0	0
F	1	1	1	0	0	-	0	1	1
G	1	1	1	0	0	0	-	1	0
H	1	1	1	1	0	1	1	-	0
I	1	1	1	1	0	1	0	0	-

(a)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	1	1	1	1	1
C	1	0	-	0	1	1	1	1	1
D	1	0	0	-	1	1	1	1	1
E	1	1	1	1	-	1	0	0	0
F	1	1	1	1	1	-	1	1	1
G	1	1	1	1	0	1	-	0	0
H	1	1	1	1	0	1	0	-	0
I	1	1	1	1	0	1	0	0	-

(b)

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
A	-	1	1	1	0	1	1	0	1
B	1	-	0	0	0	0	0	0	0
C	1	0	-	0	0	0	0	0	0
D	1	0	0	-	0	0	0	0	0
E	0	0	0	0	-	0	0	0	0
F	1	0	0	0	0	-	0	0	0
G	1	0	0	0	0	0	-	0	0
H	0	0	0	0	0	0	0	-	0
I	1	0	0	0	0	0	0	0	-

(c)

D Unit selection system at 48 kHz, using 3500 sentences

E HMM system at 48 kHz, using 500 training sentences

F HMM system at 48 kHz, using 1500 training sentences

G HMM system at 16 kHz, using 2500 training sentences

H HMM system at 32 kHz, using 2500 training sentences

I HMM system at 48 kHz, using 2500 training sentences

By comparing systems B, C and D with E, F, G, H and I, the effect of the synthesis method can be observed. By comparing systems B,C,D or G,H,I, the effect of sampling frequency, per synthesis method can be seen. Comparing systems E,F,I,the effect of the amount of training data for the HMMs is determined.

In the speaker similarity task, after the listeners listened to up to 4 original recording samples, they were presented with a synthetic speech sample generated from one of the nine systems and were asked to rate similarity to the original speaker using a 5-point scale. The scale runs from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person]. In the naturalness evaluation task, listeners used a 5-point scale from 1 [Completely Unnatural] to 5 [Completely Natural]. In the intelligibility task, the listeners heard a SUS and were asked to type in what they heard. Typographical errors and spelling mistakes were allowed for in the scoring procedure. The SUS each comprised a maximum of 6 frequently-used Romanian words.

Results

Speaker similarity – the left column of Fig. 4.7 shows the results for speaker similarity. A clear separation between the original voice (system A), HMM voices (systems E, F, G, H and I) and unit selection voices (systems B, C and D) can be initially observed. It can be also observed a clear influence of the sampling frequency over speaker similarity although improvements seem to level off at 32kHz. This is a new and interesting finding. Also there is some influence of the amount of training data. The difference between systems E and F is less significant whereas the difference between systems F and I is significant. Neither 500 nor 1500 sentences were sufficient

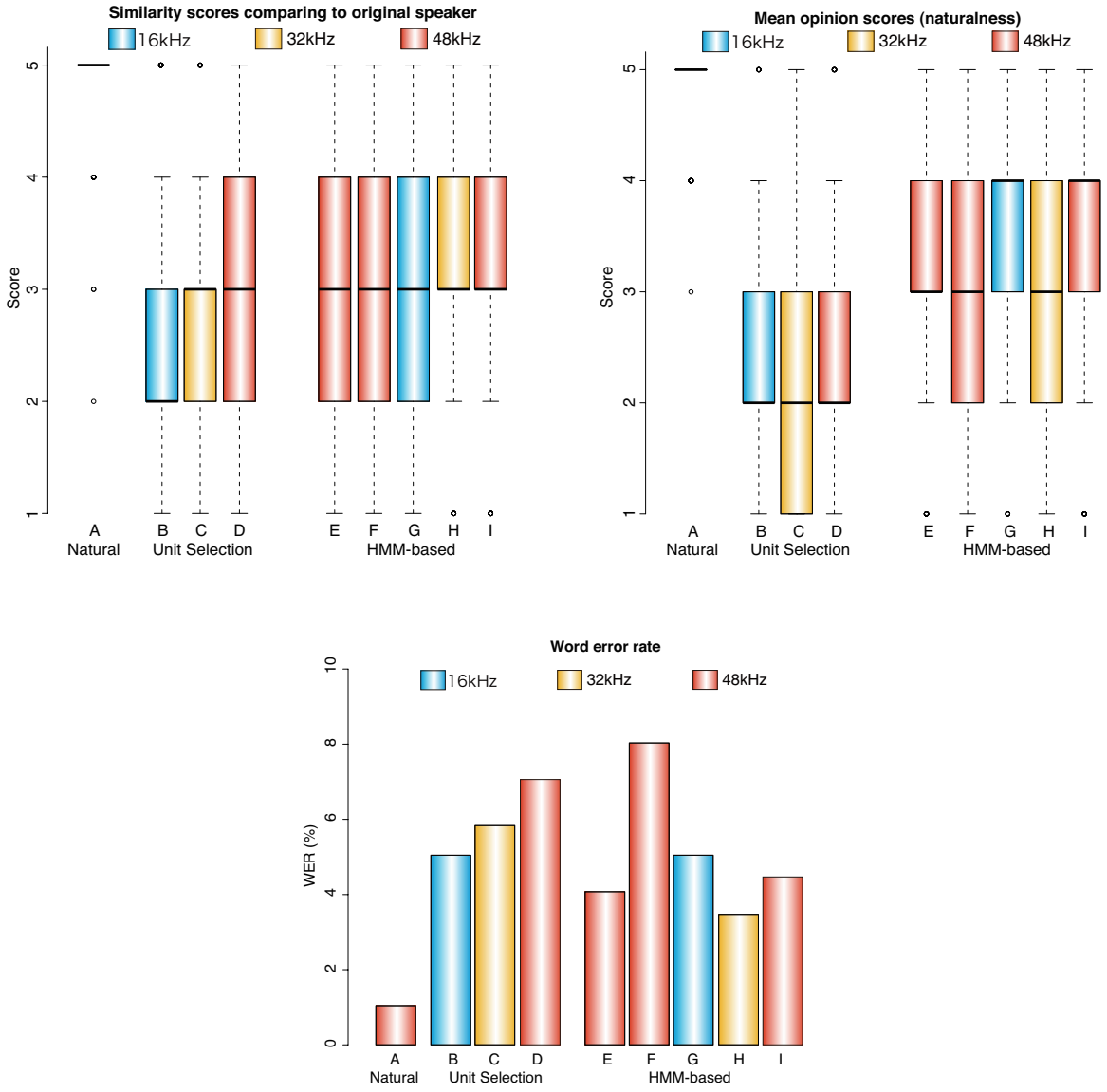


Figure 4.7: Results of the speech synthesis listening test. The graphs are box plots where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range.

for training models that can reproduce good speaker similarity, since the feature dimension is very high due to the high order mel-cepstral analysis.

Although it was expected that unit selection would have better similarity than HMM-based, the results are contrary to preliminary expectation. This may be explained by the corpus design: In the corpus, only 1000 sentences were chosen based on diphone coverage and the remaining 2500 sentences consist of 1500 randomly

chosen newspaper sentences and 1000 fairy tale sentences. Even if both types of sentences are combined, there are still 16 missing diphones and 79 diphones having fewer than 3 occurrences. Although quinphones, the base unit of HMM voices, do not have good coverage either, unit selection systems (which use diphone units) are known to be more sensitive to lack of phonetic coverage, compared to HMM-based systems [Yamagishi et al., 2008a].

Naturalness – similar tendencies to those for the similarity task can be seen, except that sampling frequency does not seem to have any effect. The use of higher sampling frequency did not improve the naturalness of synthetic speech, in contrast to speaker similarity. This is also an interesting finding. Regarding the amount of data, there are some fluctuations, although the largest amount of data typically leads to the best voice for each synthesis method.

Intelligibility –unfortunately there appears to be something of a ceiling effect on intelligibility. Absolute values of WER are generally small: both synthesis methods have good intelligibility. Even though systems D and F have a slightly higher error rate, there are no statistically significant differences between any pairs of synthetic voices in terms of WER. To confirm this a small additional test including paronyms was performed, and obtained the same results. The lack of significant differences between systems is partly caused by the nature of the simple grapheme-to-phoneme rules in Romanian. Even for SUSs and paronyms, both natural and synthetic speech are easy to transcribe, leading to WERs close to zero. This result suggests there is a need for better evaluation methods for the intelligibility of synthetic speech in languages such as Romanian.

Listening environments – to discover whether the listening environment affects the results an ANOVA test was performed. An ANOVA test at 1% significance level shows that only the system C (unit selection system at 32 kHz, using 3500 sentences) in the similarity test was affected by the listening environment. The subjects who completed the test in the listening booths generally gave lower similarity scores for system C.

Listening Test Summary

This RSS corpus is probably better suited to HMM-based synthesis than to unit selection. All speech synthesis systems built using the corpus have good intelligibility. However, a better evaluation of the system's intelligibility in simple grapheme-to-phoneme languages such as Romanian has to be designed.

The sampling frequency is an important factor for speaker similarity. More specifically, down sampling speech data in this corpus to 32kHz does no harm, but down sampling to 16kHz degrades speaker similarity. The use of higher sampling frequency, however, did not improve either the naturalness or intelligibility of synthetic speech.

These results are consistent with existing findings: [Fant, 2005] mentions that almost all the linguistic information from speech is in the frequency range 0 to 8 kHz. This implies that a 16 kHz sampling frequency (and thus 8 kHz Nyquist frequency) is sufficient to convey the linguistic information. The results also showed that using sampling frequencies over 16 kHz did not improve the intelligibility of synthetic speech. On the other hand, a classic paper regarding sampling frequency standardisation [Muraoka et al., 1978] reported that a cut-off frequency of less than 15 kHz may deteriorate audio quality. This means that the sampling frequency used should be higher than 30 kHz. In fact, the results do show that down sampling to 16kHz degrades speaker similarity. Therefore it can be concluded that the naturalness and intelligibility of synthetic speech only require transmission of linguistic information, which can be achieved at 16kHz sampling frequency, whereas speaker similarity of synthetic speech is affected by audio quality (requiring a higher sampling rate).

4.5.2 Experiment 2 – Online Interactive Demonstration

A TTS system requires an extensive evaluation of the resulted speech so that feedback from a numerous group of people can be received. The Romanian speech synthesis system is available online at `www.romaniantts.com` as an interactive demonstration. Its presence has been advertised using a series of forums with threads on Romanian TTS.

Additionally, the website includes synthesis samples in the form of the first three

chapters of the public-domain novel "Moara cu noroc" by Ioan Slavici[3] [4]. Synchronised lyrics are embedded into the mp3 files.

Since the launch of the application in August 2010, a number of 836 distinct users have accessed it. Access statistics are supplied by Google Analytics and internal IP tracking. The user input text is archived for further analysis and improvement of the synthetic speech. Analysis include: type of text, text normalisation issues and segmentation faults. Contacts have been made with users willing to use the system in their own applications, sight-impaired persons and pedagogues.

4.5.3 Experiment 3 – Adaptation to the Fairytale Speech Corpus

One of the important features of a parametric synthesiser is the possibility to adapt the parametric models to new speech corpora. [Yamagishi, 2006] describes an efficient adaptation algorithm using the Maximum Likelihood Linear Regression (MLLR) technique [Leggetter and Woodland, 1995], and Structural Maximum A Posteriori Linear Regression (SMAPLR) [Shiohan et al., 2002]. HTS had dedicated an entire unit to the speaker adaptation which can be used in conjunction to already trained models and sufficient adaptation data. It is reported that for a speaker independent or average HTS voice, as little as 5 minutes of speech are sufficient.

A preliminary adaptation experiment was carried out, trying to adapt the models of the best synthesis system to approximately 20 minutes of another person's speech. In a heuristic evaluation it was established that some of the features of the new speaker have been traced, but speaker similarity was below 3 on a MOS scale.

But in the context of the system developed, it was of great interest the adaptation of the best system[5] to the remaining fairytale subset of the RSS corpus. The idea was to enhance the pitch models by the new training data, as the fairytale subset had a more dynamic pitch domain. The 67 minutes in the fairytale data are sufficient enough, given the fact that they are reproduced by the same speaker.

After the adaptation, small increases in the intonation patterns were detected. These

[3] `http://romaniantts.com/moaraCuNoroc/chapter1-2.mp3`
[4] `http://romaniantts.com/moaraCuNoroc/chapter3.mp3`
[5] HTS, 2500 utterances, 48kHz sampling rate

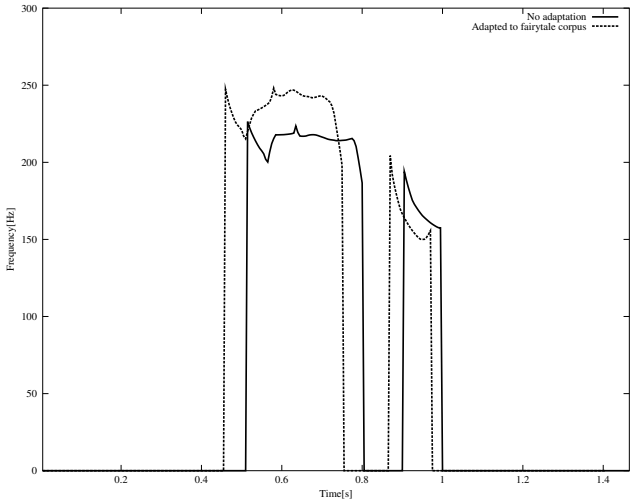


Figure 4.8: Comparison between the baseline system generated F0 contour and the contour resulted from the adaptation to the fairytale corpus

are supported by a listening test. 15 sentences were synthesised using the baseline HTS system and the adapted one. The sentences were presented for a preference evaluation to 20 listeners. The listeners used their own audio systems for the evaluation and were asked to choose the most expressive sample in each pair, using [0-No preference], [1-First sample] and [2-Second sample]. The listeners are not speech processing experts and considered the test quite hard given the subtle differences.

The results showed a 55% preference for the adapted voice, 35% for the baseline and 10% no preference. The results are thus not quite conclusive and further testing should be done. Although it is an important method to enhance already available parametric voices. Fig. 4.8 shows the difference between the original F0 contour generated by the baseline HTS system and the adapted F0 contour. It can be observed that the middle section of the utterance is slightly enhanced, and that the duration is shifted, indicating faster rhythm.

4.6 Summary

In this chapter the complete development of a Romanian HMM-based speech synthesis was described. Using the RSS corpus, some basic configuration choices made in HMM-

based speech synthesis such as the sampling frequency and auditory scale, which have been typically chosen based on experience from other fields have been revisited. The use of high sampling frequency and different auditory scales is a novel method compared to the basic HTS systems built so far.

A short theoretical overview of the HMMs and HTS was also presented along with the prerequisites for data preprocessing, such as time alignment of the HTS labels, the full set of questions for the decision trees and the speech corpus. Because the analysis of the pitch values in the fairytale corpus showed a more extended F0 domain, this subset of the RSS corpus was left aside. The questions for the decision trees include a phonetic-dependent section, that had to be rewritten according to the Romanian phoneme set.

The systems built were then compared in a Blizzard-style listening test against a minimal unit selection synthesiser and in order to determine the influence of the: sampling frequency and amount of training data. The sections included naturalness, speaker similarity and intelligibility. Higher sampling frequencies (above 16 kHz) improved speaker similarity. More specifically, the speech data in this corpus can be down sampled to 32kHz without affecting results but that down sampling to 16 kHz degrades speaker similarity. The intelligibility section arose an interesting problem: in languages with rather simple grapheme-to-phoneme rules, the errors have a ceiling effect in the sense that the systems involved in the evaluation did not have significant differences. The best HTS Romanian system was included in a online interactive demonstration at `www.romaniantts.com`.

A side evaluation was the analysis of a voice adaptation of the best HTS system to the fairytale subset, left aside from the initial training data. The results of a preference test indicated a 55% preference for the new adapted voice, but the results cannot be considered as conclusive and further evaluation should be performed, potentially using a more expressive speech corpus.

Chapter 5

A Language-Independent Intonation Modelling Technique

5.1 Introduction

The result of the speech synthesisers are often aimed more at intelligibility and less at expressivity. Especially for languages with limited resources, expressivity is hard to obtain. In order to achieve artificially enhanced expressive speech, F0 modelling techniques have been introduced. Most of the techniques use manual or semi-automated annotation of speech, which is prone to errors. Other use language-dependent phonological characteristics.

In the context of limited resources for Romanian, a language-independent model which uses no additional manual annotation would provide the much needed synthetic speech enhancements. This chapter introduces a method based on the Discrete Cosine Transform (DCT) of the phrase and syllable level of the F0 contour. Recent studies have shown that with the help of a linguistic-independent methods, more insight on the intonation effect can be derived, and contour classification based on some abstract events, other than accent, phrasing or rhythm can be achieved.

The scope of the DCT parametrisation is to use a limited, fixed, linguistic and duration independent set of parameters in the description of the F0 contour. Using no manual annotation leaves no space for subjective evaluations and thus no errors. These parameters can also be used in a statistical unsupervised classification of the intonational events.

The basic principle of the DCT parametrisation is that voiced segments of the pitch are a continuous curve that can be represented as a sum of cosine functions with different frequencies. Another important aspect of the use of DCT coefficients is the ability to model F0 at different levels or layers in accordance with an additive superpositional model –*small ripples on top of big waves* – [Sun, 2002, Sakai, 2005, Santen et al.,]. There is therefore a phrase level which determines its type (declarative, interrogative, exclamatory etc); a word level through which semantic sense can be accentuated; a syllable level or phoneme level in which the semantic emphasis relies.

Other arguments for the selection of the DCT are the use of deterministic algorithms in both the analysis and the synthesis stages of F0 modelling. The possibility of distance calculation between the parameters is also important. [Teutenberg et al., 2008] considers the ToBI model inadequate because the are no fully automated methods of annotation. Other models such as Tilt or Fujisaki do not provide the ability to measure the distance between the used characteristics and the description is based on non-deterministic occurrence of intonation events within the utterance. Also, the previous methods cannot sustain compression of data for transmission or storing.

The proposed method within this chapter is based on the DCT parametrisation of the F0 contour using two layers: phrase and syllable. The DCT coefficients are extracted using a superpositional approach and are then predicted using classification and regression trees. The results are presented in terms of correlation coefficients of the CART algorithms used, but also in terms of re-synthesis using the Romanian HTS system. The method is also evaluated so that it would be applied in an interactive intonation optimisation method using evolution strategies described in Chapter 6.

The chapter is organised as follows. Section 5.2 presents the theoretical aspects of F0 modelling, focusing on the possible problems that can occur and some of the models already proposed. From these models, the parametric DCT-based was chosen for the reasons presented above and the method is summarised in section 5.3.1. And section 5.3 describes the proposed method and its results.

5.2 F0 Modelling Techniques

This section addresses the state-of-the-art methods for F0 modelling along with their advantages and disadvantages. There is a short introduction for the need for F0 modelling continued with an overview of common techniques used. Some Romanian intonational models are also discussed.

5.2.1 Prosody

Prosody represents one of the most important areas of research for text-to-speech systems. The high quality of TTS is however limited by the use of relatively complex prosodic models, which in turn lack the naturalness of the spontaneous speech. Although the concatenation or spectrum estimation errors have been almost fully resolved, the systems have a robotic output, due to the lack of prosodic enhancement. The current tendency is to introduce several cues of prosodic information within the front-end of the system, such as semantic analysis or accent and focus emphasis, sometimes using paralinguistic information.

In linguistics, **prosody** represents the *rhythm*, *accent* and *intonation* of the speech. Prosody can reflect several different speaker characteristics, such as: emotional state, utterance type (i.e. declarative, interrogative etc.), irony and sarcasm, emphasis, contrast or focus, as well as other language dependent elements which cannot be attributed to language grammar, such as choice of words and utterance structure.

The terms of accent, rhythm and intonation belong to phonology and are associated with the cognitive aspects undertaken during speech. This means that they are abstract terms which need a physical correspondent, such as amplitude, fundamental frequency or duration [Tatham and Morton, 2005]. This correlation can be achieved through *prosody modelling*, which concerns the transformation of speech prosody to physical aspects: rise or fall of pitch, formant frequency spectrum modification or duration changes.

A basic description of prosody refers to it as the combination between **intonation** and **timing** [Taylor, 2009]. Intonation is considered as a result of the variations of the fundamental frequency, while timing refers to the duration of the speech segments. In text-to-speech systems, intonation has been the key focus of several studies and it still remains

an open issue. **Intonation** encompasses all the linguistically relevant, suprasegmental, non-lexical aspects of the fundamental frequency - or its perceptual correlate, the pitch – through the course of spoken utterances [Gronnum, 1995].

Pitch or fundamental frequency[1] is the oscillating frequency of the glottis during speech and determines its harmonic structure. It can be computed only in the voiced segments of speech and it is speaker dependent. It is influenced by the length, tension and mass of the vocal cords and the pressure of the forced expiration also called sub-glottal presure [Taylor, 2009]. The information conveyed by the pitch signal is the following: speaker gender, age and state of health, emotion, accent and prosody.

5.2.2 F0 Modelling Problems In Text-to-Speech Systems

F0 modelling represents the correlation between the speech intonation and the events within the pitch contour. But, while trying to derive a universal intonation model for TTS, several problems occur. The problems can be stated as follows:

Inter-language variability - Languages are based on specific, limited areal evolution and are a result of human interaction and mimicry. Intonation or prosodic patterns are the result of social interaction learning and have specific characteristics within a language or family of languages.

Inter-speaker variability - Speakers of the same language can sometimes express their emotions in various intonational patterns, determined by dialect or by the social upbringing.

Intra-speaker variability - Given the conversational situation, or rather the emotional state of the speaker, its intonation behaviour can vary over time.

TTS system input - the basic TTS system has as input, as its name says, the *raw text* with no additional information available. The TTS system is forced to estimate based on deterministic or probabilistic rules, a possible prosodic output. At the moment, except for some isolated cases, this output is directly related to the speech corpus used in the training section. As a paradox, most of the systems use a

[1]Pitch is considered to be the perceptual equivalent of the fundamental frequency, but in the literature they are often referred to as equivalent

rather flat intonation within the training corpus to avoid spectrum or concatenation disfluencies.

A unified model of speech production - There is no universally accepted model or method to represent or describe specific aspects of speech (e.g. accent, rhythm or intonation). Although there are quite a handful of proposed models, none of them have managed to provide a complete description of the physiological and cognitive process which result in speech, as opposed to the simplicity and robustness of the verbal components of a language (i.e. phonemes, words, phrases) [Taylor, 2009].

Semantic aspect of speech - Speech is not the result of just the chaining of several words within a sentence. It also contains the meaning, or semantics of the sentence. Different words contain a certain degree of emotion within their sense (e.g. anger within the word *hate*), so that their reproduction or reading is influenced by the underlying emotion. So far, there is no semantic memory or learning involved in a TTS, and no full semantic description yet. Some systems use punctuation, which is far from semantics, but still provides some guidelines to the outcome of the synthesis.

What is a correct prosodic model? - it is crucial to determine beforehand what the system is set to accomplish. It can be argued that the system should refrain to reading aloud a given text. On the contrary, some might wish to have a fully conversational unit, which interacts no less than a human being. It should confer the user the impression of a fully cognitive and articulatory process. Some have said that if a TTS system is to be used for example over a telephone line, the listener should not have the impression of a real person at the end of the line, as it might cause frustration. A security issue would be the replication of public personalities.

Possible Solutions

These give rise to some possible limited solutions, which can be thoroughly analysed to obtain a more elaborate, abstract model.Some of the problems above have been or are in the process of being solved using some of the following possible solutions:

- using the punctuation as a basic intonation evaluation method. Commas, dots, exclamation or question marks offer rudimentary cues to some intonational patterns.

- using prosodic annotation of the text, is a solution is some TTS systems. For example, XML[2] tags containing pitch or duration values can be used to modify the prosody. But this involves a manual annotation which is not desired in real-time applications, or for non-expert users. They cannot be extended and generalised, as it is a momentarily decision or preference.
- the use of generic semantic analysis for utterance emphasis determination
- implementing an intelligent or memory-based semantic unit for the TTS system

5.2.3 Intonation Models

The problems stated in the previous section could be solved if a correct universal F0 modelling method can be defined. The are quite a generous number of models proposed starting from phonological aspects up to simple parametrisation of the fundamental frequency as a continuous curve. The following sections present a few common F0 modelling or parametrisation techniques according to [Taylor, 2009].

The RFC and Tilt models

These models are based Palmer's idea [Palmer, 1922] to use dynamic characteristics of the F0 contour, such as *rise* and *falls*. The most important aspect of the contour is its nuclei, optionally preceded by the *head* and *pre-head* and followed by the *tail*. The prosodic events can therefore be described at nuclei level, through a rise followed by a fall[3], each with an amplitude and duration attributed (see Fig. 5.1). The main disadvantage is that these parameters cannot be interpreted and easily manipulated.

An implementation of this method is the **Tilt** model which transforms the amplitude and duration parameters into three Tilt parameters: *amplitude* is the sum of the magnitudes of the rise and fall amplitudes, *duration* is the sum of the of the rise and fall durations and *tilt* describes the global form of the international form based on its amplitude and duration. The Tilt model described the intonation as a series of events, but it does not use a fixed number of categories, but a set of continuous domain parameters.

[2]SSML - Speech Synthesis Markup Language is an XML-based markup language embeded in VoiceXML to control some aspect of the synthetic speech

[3]RFC - rise/fall/connections

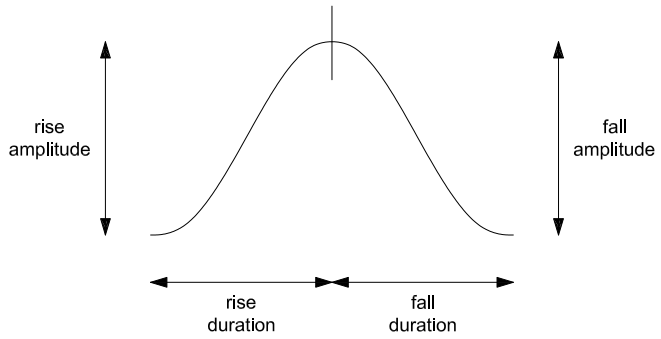


Figure 5.1: Definition of the RFC parameters

Each event has a rise and a fall, and in between there are straight lines named *connexions*. All of the variations in the accent form and the edge form are obtained through the modification of the rise and fall components, as well as the way in which the events are aligned with the speech components.

Autosegmental-Metrical and ToBI models

These are one of the most popular F0 modelling models. The Autosegmental-Metrical (AM) or Tones and Break Indices (ToBI) model describes the pitch through a series of low (L) and high (H) tones in combination with a series of diacritics which differentiate between the tones located on accented syllables from those occurring at boundaries and between accents. Accents can be composed of one or two tones. Each accent can be directly associated with the accented syllable, noted by "*", or can be a boundary tone, marked by a "%", or a phrase tone, marked by a "-". The list of possible pitch accents is: H*, L*, H* + L, H + L*, L + H* and L* + H. An example of a simple TOBI annotation is shown in Fig. 5.2. The break indices mark the the boundary strength between adjacent words on a scale of [0 - No boundary] to [4 - Phrase boundary].

As opposed to the British school, there is no separation between the nuclei and the head for example. Each of these events can be described by any of the values presented above. Each tone represents a target and the pitch values are obtained through interpolation. Even the HTS labels use this type of annotation, but still require a manual annotation,

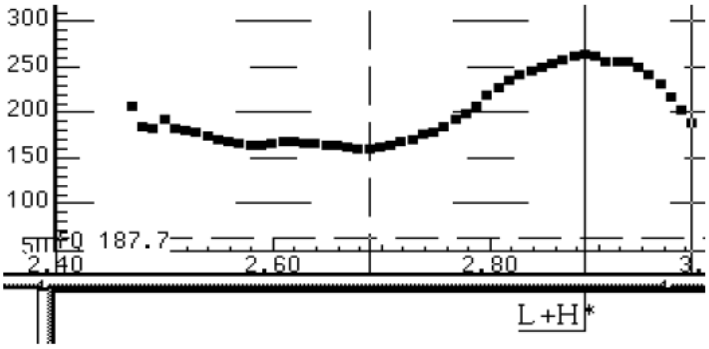


Figure 5.2: Example of ToBI annotation (from [Jilka et al., 1999])

which reduces their use and applicability. It is important to note that ToBI is a labelling system and does not specify the means to produce quantitative intonation from the ToBI labels [Benesty et al., 2007].

The INTSINT model

This model was developed in order to allow a multi-lingual annotation of the intonation. In [Hirst and Cristo, 1998] the full model and the method in which it can be applied to multiple languages is described. The principle is the description of a series of prosodic events through a limited number of attributes with a rather abstract character, so that they are not language or speaker dependent. A set of target points are used, which are defined relative to the speaker's fundamental frequency or relative to the previous target point.

The Fujisaki Model

The Fujisaki model tries to provide an accurate description of the F0 contour based on the way in which it is produced in speech. The model has two components: *phrase* and *accent*. It is a derivation of the filter method proposed by Ohman [Ohman, 1967]. The input of the model are impulses used to produce phrase shapes and step functions for the accent shapes. The implementation contains two second order FIR filters, one for each component. It is considered to be the first superpositional model for F0 modelling.

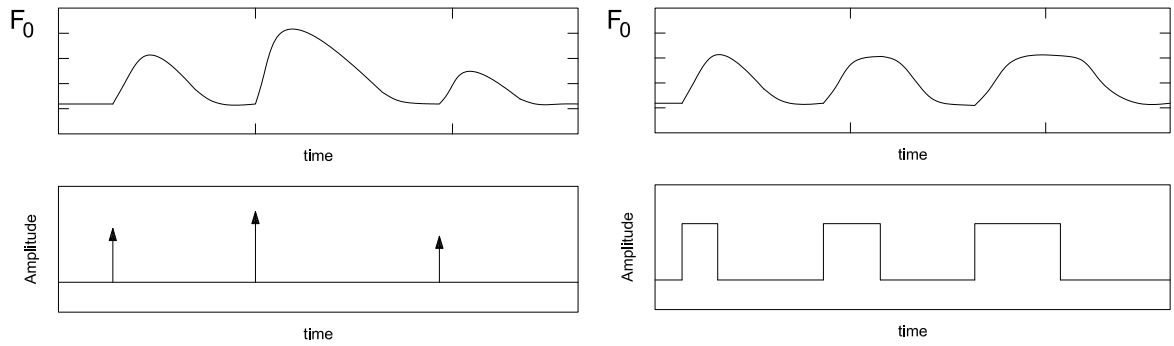


Figure 5.3: Example of phrase (a) and accent (b) excitations and contours used by the Fujisaki model

Comparison

Table 5.1 presents a comparison between the methods presented above in terms of purpose of the model, type (phonological, phonetic or acoustic), character of pitch representation (tone or pitch shapes and dynamics) and the number of levels of parametrisation (superpositional or linear).

Table 5.1: A comparison between most common F0 modelling techniques

Model		
Tilt	Purpose	Description of pitch events
	Type	Acoustic
	Pitch representation	Shape
	Level	Linear
ToBI	Purpose	Theory of how intonation occurs in human communication
	Type	Phonological
	Pitch representation	Tone
	Level	Linear
INTSINT	Purpose	An equivalent to IPA for intonation
	Type	Phonetic
	Pitch representation	Shape
	Level	Linear
Fujisaki	Purpose	Reproduction of the actual articulation
	Type	Acoustic
	Pitch representation	Shape
	Level	Superpositional

Romanian Intonation Models

For Romanian both deterministic and statistical models of intonation have been proposed. An important resource is [Hirst and Cristo, 1998], where D. Jinga defines a series of intonational patterns using simple semantics and syntactics. In his work, [Bodo, 2009] applies these studies to a diphone concatenation synthesis system, with satisfactory results. A more basic approach is presented in [Ferencz, 1997], where prosody can be manually manipulated through a number of parameters set by the user.

The works of [Apopei and Jitcă 2008], [Jitcă et al., 2008], [Apopei and Jitcă 2006], [Apopei and Jitcă 2007] and [Apopei and Jitcă 2005] define a set of intonational categories for Romanian, and the methods through which these categories can be directly derived from text with additional tags.The categories are used in a formant synthesis system.

A series of analysis of the accent influence over the word and phrase level F0 contour, as well as speech rhythm and speed are presented in [Giurgiu, 2006, Giurgiu and Peev, 2006]. An interesting approach using the Topic-Focus Articulation method has been used in [Curteanu et al., 2007].

5.2.4 Discrete Cosine Transform

DCT is a discrete transform which expresses a sequence of discrete points as a sum of cosine functions oscillating at different frequencies with zero phase. The are several forms, but the most common one is DCT-II. The coefficients are computed according to Eq. 5.1.

$$C(u) = \alpha(u) \sum_{x=0}^{N-1} f(x) cos \left[\frac{\pi(2x+1)u}{2N} \right], u = 0, 1, 2, ..., N-1 \tag{5.1}$$

The inverse transform of the DCT is presented in Eq. 5.2.

$$f(x) = \sum_{u=0}^{N-1} \alpha(u) C(u) cos \left[\frac{\pi(2x+1)u}{2N} \right], x = 0, 1, 2, ..., N-1 \tag{5.2}$$

In both equations $\alpha(u)$ is defined as:

$$\alpha(u) = \begin{cases} \sqrt{1/N} & , u = 0 \\ sqrt(2/N) & , u \neq 0 \end{cases} \tag{5.3}$$

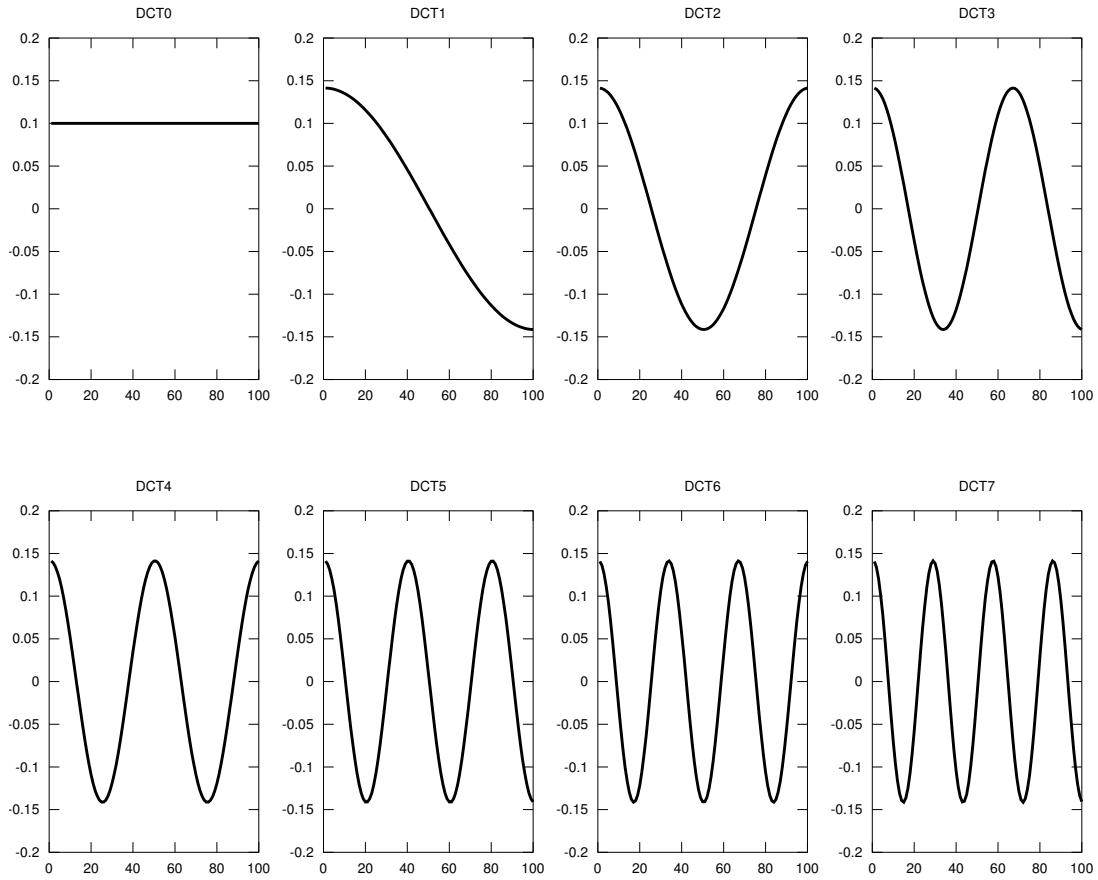


Figure 5.4: The first 8 DCT basis cosine functions

Discrete Cosine Transform is commonly used in signal and image processing. Its advantages rely in the fact that it can compress and concentrate most of the energy of a signal into relatively few low-frequency coefficients. DCT is also a good approximation to principal component analysis, also known as Karhunen-Loeve transform [Vaseghi, 2007].

DCTs are also used for solving partial differential equations by spectral methods and Chebyshev polynomials. Fast DCT algorithms are used in Chebyshev approximation of arbitrary functions by series of Chebyshev polynomials, for example in Clenshaw-Curtis quadrature.

The first 8 DCT basis cosine functions are represented in Fig. 5.4.

5.3 F0 Parametrisation using the Discrete Cosine Transform

Until recent, intonation was considered to be strictly part of the underlying phoneme, with little or no regard towards the effects of the context and longer states such as syllables or words. Some recent studies [Sun, 2002] define the intonation as a superpositional model of several overlapping layers. Fujisaki's [Fujisaki and Ohno, 1998] intonation model is in some ways an implementation of the superpositional principle, in the sense that the accent components are overlapped to the phrase component. As opposed to Fujisaki's model, recent studies introduced superior linguistic levels, such as syllables or words.

5.3.1 Related Work

In the last few years, a series of articles have addressed the issue of DCT parametrisation of the F0 contour. The first work to use DCT as a pitch parametrisation method is presented by [Muralishankar et al., 2004]. The authors use the DCT coefficients to modify the pitch, based on the linear prediction residual. [Teutenberg et al., 2008] is the first work to use DCT coefficients for the prediction of the F0 contour for speech synthesis. It uses two pitch layers: phrase and voiced segment. The phrase level is represented by a contour that passes through the mean pitch value of every voiced segment. This means that the first coefficient of the voiced segments can be excluded. Because the unvoiced segments of the pitch have undefined F0 values, these segments are linearly interpolated. At the voiced segment level, DCT coefficients are extracted with no additional processing.

To determine the correct number of DCT coefficients which does not introduce a significant error, an analysis was performed. The F0 contour is parametrised using the discrete cosine transform and then the number of coefficients is limited. The results indicate a value of 6 coefficients for the phrase level and 10 for the voiced segments.

For the synthesis stage, classification and regression trees are built for each coefficient. The parameters used in the nodes of the tree are: segment duration, accent position, segments position for the voiced segments and phrase duration, accent position and phrase position for the phrase level. The errors are reported to be smaller than in other methods, such as [Sakai, 2005, Sun, 2002].

[Latorre and Akamine, 2008] continues the ideas in [Teutenberg et al., 2008], but adds a series of conditions and parameters. Instead of the voiced segments, the syllable level is introduced. It also uses the parametrisation of the $logF0$ syllable level contour. Linear interpolation is also used in the unvoiced segments of the F0. To limit the interpolation errors, pitch values were limited using the following conditions: F0 values have an auto-correlation coefficient higher than 80%; they belong to the phonemes which have a clear periodicity (vowels, semivowels, nasals) and have a value that does not exceed a margin of half an octave around the average value of the syllable's logF0.

Decision trees are used to group the parameters. Similar models are built for other linguistic levels. For the synthesis, the statistical model of each layer is used to define a log-likelihood function. These models are weighted and summed into a global log-likelihood which is maximised in respect to the syllable level DCT coefficients. The tests shown a clear preference towards the proposed model against the pure HMM model. For the training and prediction stages, several parameters have been added. These parameters are classified into concatenation and description parameters. The concatenation ones define the neighbouring syllable, through first coefficient delta parameter ($\Delta DCT0$) and the F0 gradient in the concatenation point between syllables ($\Delta LogF0_s^{prev}$, $\Delta LogF0_s^{next}$). The description parameters refer to the current syllable F0 contours through logF0 variance ($Var(logF0)$). These four parameters along with the 7 DCT coefficients used at syllable level determine decisions in the classification and regression trees. Their results conclude to a more dynamic F0 contour, considered more natural.

Another work which used the DCT parametrisation of the F0 contour is [Wu et al., 2008]. The work used the phrase and syllable level. As a novelty, there is no interpolation of the unvoiced segments. To estimate the DCT coefficients, classification and regression trees based on Maximum Likelihood (ML) and Minimum Description Length (MDL) are used. Along with the DCT parametrisation, Natural Cubic Spline (NCS) is evaluated. NCS determined higher errors in the prediction stage. The Robust Algorithm for Pitch Tracking (RAPT) extract the F0 values. They use 7 DCT coefficients, delta and delta-delta parameters of the first DCT coefficient for the syllable level. For the phrase level the first 3 DCT coefficients were selected to represent a contour that passes through each syllable's mean value.

The most recent and elaborate work in this area is [Qian et al., 2009]. It continues the work of [Wu et al., 2008], but additionally addresses the issue of the acoustic segments duration. The F0 extraction method and parameters are the same as in [Wu et al., 2008]. The differences rely in the use of a state level contour, generated by the Markov chains and the maximisation of the joint probability of state and higher levels.

5.3.2 Proposed Method

The works of [Latorre and Akamine, 2008], [Qian et al., 2009], [Teutenberg et al., 2008] and [Wu et al., 2008] presented the arguments and results of DCT parametrisation of the pitch contour. These methods are all based on the superpositional principle of intonation. The results presented in the articles indicate a justification for this method of parametrisation.

Starting from these studies, a new method for F0 parametrisation using DCT is proposed [Stan and Giurgiu, 2011]. The novelty relies in the parametrisation of the syllable level, that is, the IDCT of the phrase level coefficients are subtracted from the original F0 contour[4]. In other words, the syllable is considered to have an additive effect over the phrase level and it is not considered as absolute value. On the other hand, in [Qian et al., 2009] the authors mention the use of just the vocal segments for parametrisation, although it is not clear how they evaluated the DCT coefficients of just these segments. Therefore the implemented method uses a linear interpolation of the unvoiced segments. The number of DCT coefficients is **8** for the phrase level and **7** for the syllable level. This choice is based on a preliminary evaluation of the error introduced by the limiting of the DCT coefficients. For the phrase level, 8 coefficients were selected, because DCT0 represents the mean of the curve and it is considered to be speaker dependent.

Fig. 5.5 presents the error values with respect to the number of the DCT coefficients used to parametrise a random syllable contour. The number of coefficients varies from 1 to the length of the syllable. It can be observed that for a number as low as 5 DCT coefficients, the error is around 15Hz.

[4]It might be argued that the syllable level includes the phoneme level as well, but using just a limited number of coefficients for the syllable level, the phoneme level is considered to be the remaining F0 variation

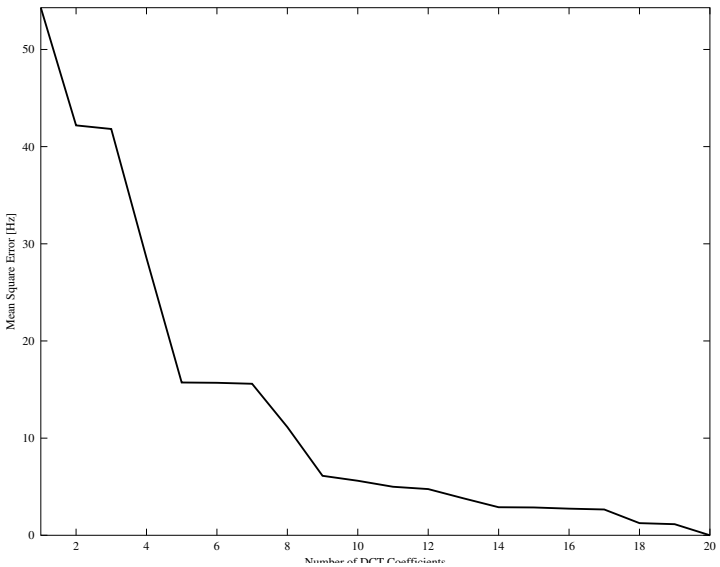


Figure 5.5: Error of the DCT coefficients truncation in the prediction of a random syllable F0 contour.

The **steps** of the method are the following:

(1) Extraction of the F0 contour from the entire audio corpus using a voting method between: Instantaneous Frequency Amplitude Spectrum (IFAS), Fixed Point Analysis and Entropic Signal Processing system (ESPS)

(2) Linear interpolation of the unvoiced segments

(3) Phrase level segmentation

(4) Extraction of the first 8 DCT coefficients from the phrase level contour

(5) Subtracting the IDCT of the phrase level contour from the original one

(6) Syllable level segmentation

(7) Extraction of the first 7 DCT coefficients from the syllable level contour

(8) Additional feature extraction

(9) Classification and regression tree training

(10) DCT coefficients prediction based on the best algorithm selected in the training stage

(11) Comparing the predicted F0 contours with the ones generated by the baseline synthesis system

(12) Speech synthesis using the predicted F0 contour and audio evaluation of the result

5.3.3 Audio Corpus Preprocessing

In order to estimate the DCT coefficients of the F0 contour using classification and regression trees, a training data set is needed. A subset of the RSS corpus was selected, namely *rnd1*, a 500 random selected newspaper utterances. The corresponding HTS labels were also used. The phrase and syllable level segmentation was achieved using these labels.

After the segmentation a number of **730 phrases** and **13029** syllables were identified within the *rnd1* subset. For the evaluation part, the first 10 utterances were set aside. They contain **16 phrases** and **301** syllables. The DCT coefficients were extracted using self-implemented Python scripts. Along with the DCT coefficients for each level, a series of parameters were added to the feature vectors, as follows:

Phrase level:

- number of syllables in {previous, current, next} phrase
- number of words in {previous, current, next} phrase
- position of the current phrase in utterance {forward, backward}
- number of syllables in utterance
- number of words in utterance
- number of phrases in utterance
- length of phrase expressed in F0 samples - the sampling period is 5 ms.

So that the feature vector at phrase level is composed of the features above plus the first **8** DCT coefficients at phrase level, a total of **20 parameters**.

Syllable level:

- accent of the {previous, current, next} syllable
- number of phonemes in {previous, current, next} syllable
- number of syllables in {previous, current, next} phrase
- number of words in {previous, current, next} phrase
- position of current syllable in the current word {forward, backward}
- position of current syllable in the current phrase {forward, backward}

- number of accented syllables before current syllable in the current phrase
- number of accented syllables after the current syllable in the current phrase
- the number of accented syllables from the previous accented syllable to the current syllable
- the number of accented syllables from the current syllable to the next accented syllable
- name of the vowel of the current syllable
- position of the current phrase in the utterance {forward, backward}
- number of syllables in utterance
- number of words in utterance
- number of phrases in utterance
- length of syllable expressed in F0 samples - the sampling period is 5 ms.

The syllable level feature vector therefore comprises a number of **40 parameters**, i.e. **7** DCT coefficients and the features presented above.

5.3.4 Attribute Selection

The number of features contained in the HTS labels is rather high, so that using the capabilities of the Weka[5] tool, a preliminary attribute selection step was implemented. Weka is a collection of algorithms for machine learning and data mining. For the attribute selection, a greedy stepwise without backtracking algorithm was selected. Greedy stepwise can either move forward or backward within the search space and select the best feature using cross validation. The results indicate the use of the following sets of attributes:

Phrase Level	
DCT0	number of syllables in previous phrase
	number of words in previous phrase
	number of syllables in current phrase

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

	number of words in current phrase
	position of current phrase in utterance forward
	number of phrases in utterance
	length of phrase expressed in F0 samples
DCT1	position of the phrase in utterance backward
	length of phrase expressed in F0 samples
DCT2	position of the phrase in utterance backward
DCT3	number of syllables in current phrase
	position of current phrase in utterance backward
	number of syllables in current word
DCT4	position of the phrase in utterance backward
	length of phrase expressed in F0 samples
DCT5	position of the phrase in utterance forward
	number of phrases in utterance
DCT6	number of syllables in next utterance
	number of syllables in previous utterance
DCT7	position of the phrase in utterance forward
	number of phrases in utterance

Syllable Level	
DCT0	accent of the previous syllable
	number of phonemes in previous syllable
	position of the current syllable in the current word forward
	position of the current syllable in the current word backward
	number of previous accented syllables in the current phrase
	number of next accented syllables in the current phrase
	number of syllables from the previous accented syllable
	the name of the vowel in the current syllable
DCT1	accent of the current syllable
	number of phonemes in current syllable

	position of the current syllable in the current word forward
	the name of the vowel in the current syllable
	position of the current phrase in utterance
DCT2	accent of the previous syllable
	number of phonemes in previous syllable
	accent of the current syllable
	the name of the vowel in the current syllable
	accent of the next syllable
	number of phonemes in next syllable
	number of phrases in utterance
DCT3	number of phonemes in previous syllable
	number of phonemes in current syllable
	the name of the vowel in the current syllable
	accent of the next syllable
	number of phrases in utterance
DCT4	number of phonemes in previous syllable
	position of the current syllable in the current word forward
	position of the current syllable in the current word backward
	number of syllables to next accented syllable
	the name of the vowel in the current syllable
	number of phonemes in next syllable
	position of the current phrase in utterance backward
	number of words in next phrase
	number of phrases in utterance
	length of the syllable expressed in F0 samples
DCT5	accent of the previous syllable
	number of phonemes in previous syllable
	accent of the current syllable

	position of the current syllable in current word forward
	the name of the vowel in the current syllable
	number of phonemes in next syllable
	position of the phrase in the current utterance backward
	length of the syllable expressed in F0 samples
DCT6	number of phonemes in previous syllable
	position of the current syllable in the current word forward
	number of syllables to the next accented syllable
	the name of the vowel in the current syllable
	accent of the next syllable
	number of phonemes in the next syllable
	position of the current syllable in the current phrase
	number of syllables in the next phrase
	length of the syllable expressed in F0 samples

The features presented above will be called *the reduced set of features*. A later evaluation estimated the performance trade using this reduced set instead of the full feature set.

For a correct implementation of the DCT-based method, preliminary statistics of the DCT coefficients are in order. The histograms of the phrase and syllable level DCT coefficients of the *rnd1* subset are shown in Fig. 5.7 and 5.6. Tables 5.3 and 5.4 present the statistics of these histograms in terms of mean, standard deviation, minimum and maximum values. And Table. 5.2 shows the average duration of pitch contour in each of the two levels used.

Table 5.2: Statistics of the *rnd1* subset phrase and syllable lengths, given in seconds.

	Mean	**Std. dev.**	**Min**	Max
Syllable	0.150	0.068	0.025	0.763
Phrase	1.694	3.161	0.319	8.265

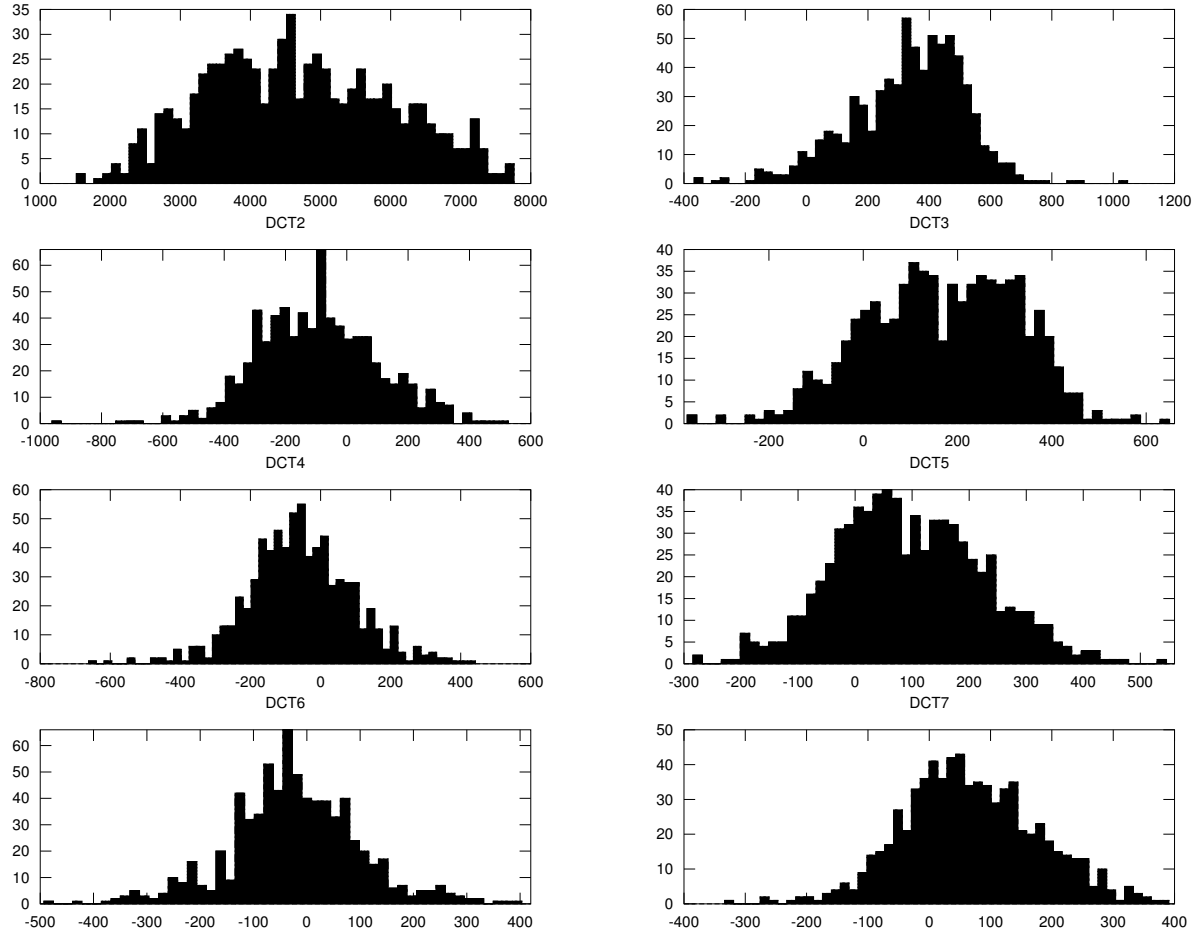


Figure 5.6: **Phrase** level DCT coefficients histograms

Table 5.3: Statistics of the **phrase** level DCT coefficients. 730 coefficients were analysed corresponding to the number of phrases in *rnd1*

DCT Coefficient	Mean	Std. dev.	Min.	Max.
DCT0	4690.300	1318.300	1511.162	7762.336
DCT1	331.750	185.850	-366.800	1046.777
DCT2	-95.087	197.470	-961.830	526.653
DCT3	168.270	161.030	-314.262	652.300
DCT4	-57.100	151.600	-787.123	446.700
DCT5	94.427	130.15 0	-298.882	552.150
DCT6	-22.312	123.020	-501.100	409.565
DCT7	67.095	110.370	-335.890	390.000

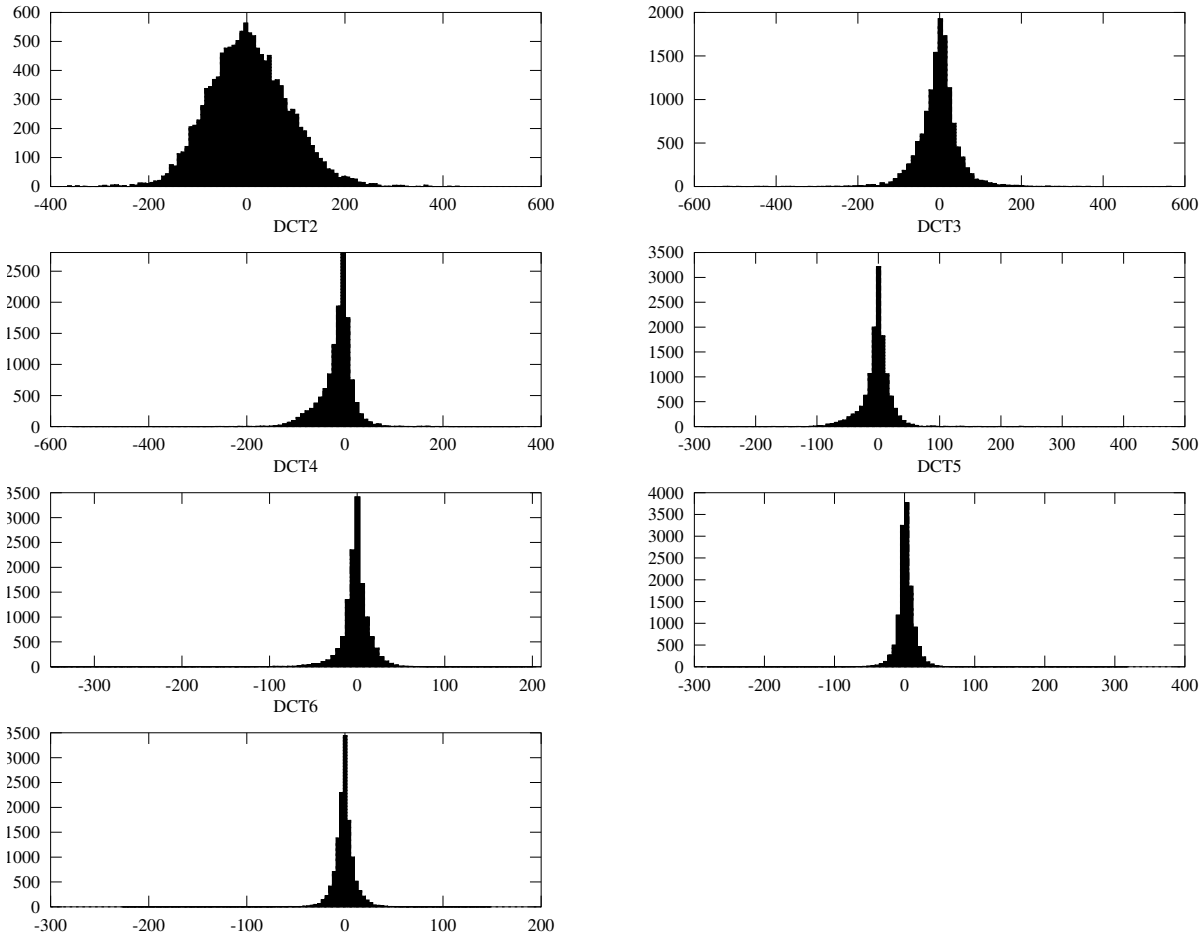


Figure 5.7: **Syllable** level DCT coefficients histograms

Table 5.4: Statistics of the **syllable** level DCT coefficients. 13029 coefficients were analysed corresponding to the number of syllables in *rnd1*

DCT coefficient	Mean	Std. dev.	Min.	Max.
DCT0	33.82	82.21	-365.68	433.13
DCT1	18.18	50.31	-528.90	566.26
DCT2	-99.45	34.84	-555.98	357.08
DCT3	58.54	25.63	-284.07	401.15
DCT4	-74.09	19.14	-349.23	201.05
DCT5	1.93	15.01	-225.75	317.72
DCT6	-0.04	12.96	-235.32	202.99

5.4 Evaluation

5.4.1 Experiment 1 – CART Training

Using the features described in sections 5.3.3 and 5.3.4 the Attribute-Relation File Format (ARFF) files used for the classification and regression tree training and testing were built. ARFF is Weka compliant and it is a simple easily-readable file format.

A first step of the training refers to the determination of the best training and prediction algorithm for the data set used. A selection of the algorithms presented in [Witten and Frank, 2005] based on speed and efficiency was made. These include:

Linear Regression - is an approach to modelling the relationship between a scalar variable and one or more input variables using linear functions.

M5 Rules - is an algorithm deriving regression rules from the model trees built using M5. M5 is a decision tree which contains linear models in its leafs.

Additive Regression - is a method based on weighted sums of trees.Several trees are built using a CART principle (for example M5) and their leaf nodes are weighted into a new sum function.

Each of these algorithms were evaluated for both phrase and syllable level. In order to evaluate the correct attribute selection previously described, the results of the reduced set of features was also compared to the results of the full set. A separate tree is built for each DCT coefficient. The results presented in tables 5.5, 5.6, 5.7 and 5.8 were obtained using a 10-fold cross validation.

It can be observed that the reduced set of features has similar performances as the full set of features. Based on the results presented, the Additive Regression algorithm was selected for the prediction of all of the phrase level DCT coefficients. The same algorithm was used for the syllable level DCT coefficients. Because of the small differences between the full and reduced sets of features and in order to reduce complexity and computational costs, the reduced set was selected for the prediction stage.

Table 5.5: Results of the **phrase** level DCT coefficients prediction using the **full set** of features. Columns in the table represent: [1] Correlation Coefficient, [2] Mean Absolute Error, [3] Root Mean Squared Error, [4] Relative Absolute Error [%], [5] Root Relative Squared Error [%].

DCT Coefficient	**Algorithm**	[1]	[2]	[3]	[4]	[5]
	Linear Regression	0.97	52.94	69.42	19.14	20.51
DCT0	M5 Rules	0.96	280.65	379.53	25.72	28.83
	Additive Regression	0.98	175.79	220.96	16.11	16.78
	Linear Regression	0.57	116.33	152.91	79.43	81.74
DCT1	M5 Rules	0.62	110.54	146.20	75.48	78.15
	Additive Regression	0.64	106.79	143.46	72.92	76.68
	Linear Regression	0.53	126.49	167.28	80.82	84.41
DCT2	M5 Rules	0.65	111.58	150.25	71.29	75.88
	Additive Regression	0.66	109.23	147.89	69.79	74.69
	Linear Regression	0.47	114.61	141.66	85.95	87.93
DCT3	M5 Rules	0.59	102.89	129.73	77.16	80.52
	Additive Regression	0.59	102.90	129.77	77.18	80.55
	Linear Regression	0.40	106.83	138.83	91.73	91.57
DCT4	M5 Rules	0.52	99.03	129.00	85.03	85.09
	Additive Regression	0.53	97.25	128.02	83.50	84.45
	Linear Regression	0.34	96.02	122.00	91.06	93.60
DCT5	M5 Rules	0.39	93.30	119.72	88.48	91.86
	Additive Regression	0.42	91.71	117.87	86.98	90.44
	Linear Regression	0.29	89.28	117.53	96.12	95.47
DCT6	M5 Rules	0.42	84.79	111.59	91.29	90.64
	Additive Regression	0.45	82.62	109.42	88.96	88.88
	Linear Regression	0.20	85.61	108.15	97.64	97.99
DCT7	M5 Rules	0.21	85.21	109.02	97.18	98.77
	Additive Regression	0.24	84.37	107.74	96.22	97.62

Table 5.6: Results of the **phrase** level DCT coefficients prediction using the **reduced set** of features. Columns in the table represent: [1] Correlation Coefficient, [2] Mean Absolute Error, [3] Root Mean Squared Error, [4] Relative Absolute Error [%], [5] Root Relative Squared Error [%].

DCT Coefficient	Algorithm	[1]	[2]	[3]	[4]	[5]
DCT0	Linear Regression	0.97	202.56	262.74	18.56	19.96
	M5 Rules	0.96	275.77	375.04	25.27	28.49
	Additive Regression	0.98	179.20	223.73	16.42	16.99
DCT1	Linear Regression	0.54	119.99	156.43	81.93	83.62
	M5 Rules	0.62	111.36	146.17	76.04	78.13
	Additive Regression	0.63	109.60	144.60	74.84	77.30
DCT2	Linear Regression	0.53	128.75	167.06	82.26	84.37
	M5 Rules	0.55	127.80	164.74	81.65	83.20
	Additive Regression	0.54	129.31	166.37	82.62	84.02
DCT3	Linear Regression	0.44	110.23	171.60	88.95	87.93
	M5 Rules	0.53	112.90	135.37	77.16	82.26
	Additive Regression	0.57	102.90	129.77	78.81	81.55
DCT4	Linear Regression	0.38	116.38	140.33	93.73	95.70
	M5 Rules	0.52	97.25	132.20	84.25	86.42
	Additive Regression	0.53	101.30	131.10	84.03	86.09
DCT5	Linear Regression	0.32	97.22	125.33	92.06	96.60
	M5 Rules	0.37	92.71	117.87	86.98	91.24
	Additive Regression	0.39	93.30	120.72	89.48	92.86
DCT6	Linear Regression	0.27	90.82	118.53	97.21	96.27
	M5 Rules	0.42	83.22	110.22	89.96	92.88
	Additive Regression	0.42	85.97	111.77	92.29	91.64
DCT7	Linear Regression	0.19	87.23	110.15	98.64	97.99
	M5 Rules	0.20	85.37	110.40	96.22	97.62
	Additive Regression	0.21	86.21	109.30	97.60	99.77

Table 5.7: Results of the **syllable** level DCT coefficients prediction using the **full set** of features. Columns in the table represent: [1] Correlation Coefficient, [2] Mean Absolute Error, [3] Root Mean Squared Error, [4] Relative Absolute Error [%], [5] Root Relative Squared Error [%].

DCT Coefficient	**Algorithm**	[1]	[2]	[3]	[4]	[5]
DCT0	Linear Regression	0.40	58.01	75.24	89.61	91.52
	M5 Rules	0.49	54.72	71.56	84.52	87.04
	Additive Regression	0.59	50.57	66.30	78.11	80.64
DCT1	Linear Regression	0.23	31.10	48.09	96.84	97.28
	M5 Rules	0.34	29.94	46.95	93.23	94.98
	Additive Regression	0.37	29.47	46.18	91.76	93.42
DCT2	Linear Regression	0.25	22.86	34.14	96.22	96.81
	M5 Rules	0.28	22.46	33.82	94.53	95.91
	Additive Regression	0.30	22.36	33.66	94.14	95.45
DCT3	Linear Regression	0.26	15.07	24.95	98.11	96.57
	M5 Rules	0.26	15.05	25.38	98.02	98.23
	Additive Regression	0.28	15.04	25.16	97.94	97.39
DCT4	Linear Regression	0.19	11.13	18.82	102.74	98.10
	M5 Rules	0.23	10.89	18.89	100.49	98.49
	Additive Regression	0.24	10.89	18.86	100.48	98.03
DCT5	Linear Regression	0.05	8.72	15.00	99.84	99.91
	M5 Rules	0.02	8.79	15.33	100.65	102.11
	Additive Regression	0.05	8.73	15.04	100.01	100.20
DCT6	Linear Regression	0.08	7.55	12.92	100.81	99.69
	M5 Rules	0.10	7.58	13.05	101.26	100.69
	Additive Regression	0.05	7.54	13.06	100.74	100.73

Table 5.8: Results of the **syllable** level DCT coefficients prediction using the **reduced set** of features. Columns in the table represent: [1] Correlation Coefficient, [2] Mean Absolute Error, [3] Root Mean Squared Error, [4] Relative Absolute Error [%], [5] Root Relative Squared Error [%].

DCT Coefficient	Algorithm	[1]	[2]	[3]	[4]	[5]
DCT0	Linear Regression	0.37	58.67	76.24	90.62	92.73
	M5 Rules	0.49	54.72	71.56	84.52	87.04
	Additive Regression	0.52	53.55	70.15	82.72	85.32
DCT1	Linear Regression	0.20	30.80	48.35	95.92	97.81
	M5 Rules	0.21	30.68	48.29	95.55	97.69
	Additive Regression	0.21	30.61	48.24	95.32	97.59
DCT2	Linear Regression	0.17	23.23	34.73	97.79	98.46
	M5 Rules	0.16	23.23	34.74	97.81	97.86
	Additive Regression	0.17	23.23	34.73	97.79	98.46
DCT3	Linear Regression	0.25	14.89	24.99	96.97	96.74
	M5 Rules	0.26	14.84	24.93	96.68	96.50
	Additive Regression	0.26	14.84	24.93	96.65	96.48
DCT4	Linear Regression	0.15	10.88	18.95	100.44	98.83
	M5 Rules	0.18	10.97	18.91	101.21	98.58
	Additive Regression	0.19	10.95	18.86	101.06	98.32
DCT5	Linear Regression	0.05	8.71	15.00	99.68	99.88
	M5 Rules	0.06	8.71	15.00	99.77	99.90
	Additive Regression	0.06	8.72	15.02	99.87	100.03
DCT6	Linear Regression	0.08	7.52	12.92	100.50	99.65
	M5 Rules	0.07	7.56	12.93	101.01	99.79
	Additive Regression	0.10	7.53	12.90	100.45	99.53

5.4.2 Experiment 2 – DCT Coefficients Prediction using Additive Regression

All of the data preprocessing and training has been done in order to prepare a prediction model for the DCT coefficients using only the features available in HTS labels, even more, the reduced set of features. The DCT coefficients were extracted from the linear interpolated F0 contours of the *rnd1* subset of the RSS corpus at phrase and syllable level. Classification and regression trees were then trained for each of the coefficients separately. Using the best algorithm determined at this stage, the prediction of the DCT coefficients from the testing set of 10 utterances was performed. Table 5.9 presents the estimation error for each of the 15 DCT coefficients. It can be observed that the higher the order of the coefficient, the higher the error. This is due to the wider standard deviation and less correlation factor between the features used and the coefficients.

Table 5.9: Results of the DCT coefficients **prediction** using the Additive Regression algorithm. Columns in the table represent: [1] Correlation Coefficient, [2] Mean Absolute Error, [3] Root Mean Squared Error, [4] Relative Absolute Error [%], [5] Root Relative Squared Error [%].

	DCT Coefficient	[1]	[2]	[3]	[4]	[5]
Phrase	DCT0	0.99	108.99	133.89	8.54	9.29
	DCT1	0.57	104.64	119.99	105.10	92.78
	DCT2	0.64	114.48	141.04	93.14	80.71
	DCT3	0.59	78.77	97.60	71.63	75.76
	DCT4	0.59	81.83	96.40	78.90	72.59
	DCT5	0.52	61.79	86.01	75.74	89.91
	DCT6	0.63	43.04	53.32	88.98	83.99
	DCT7	0.72	46.80	71.69	61.89	72.72
Syllable	DCT0	0.66	38.12	49.57	74.36	76.60
	DCT1	0.45	24.26	36.13	87.96	93.79
	DCT2	0.28	20.16	29.88	94.26	96.86
	DCT3	0.36	13.03	19.58	94.90	94.78
	DCT4	0.14	10.02	15.24	104.08	104.78
	DCT5	0.15	18.73	15.04	100.00	100.20
	DCT6	0.15	7.54	13.06	100.74	100.73

5.4.3 Experiment 3 – Listening Test

A secondary experiment was conducted in order to determine the perceivable error of the F0 contour estimation. It involved the synthesis of the test sentences using the benchmark HTS system generated F0 contour versus the predicted F0 contour. Figure 4 and Figure 5 show a comparison between these contours. Although it can be easily observed that higher variations in the F0 contour cannot be followed with accuracy by the predicted coefficients, the mean error for the global F0 contour is 15 Hz, which is comparable to the one obtained by [Latorre and Akamine, 2008] or [Sun, 2002], 13Hz for the syllable level and 8 Hz for the phrase level. The higher error value for the syllable level supports the idea of introducing a separate level for phonemes as future work. A small listening test was also set up, 10 listeners were presented with 20 pairs of utterances consisting of the baseline synthesiser and the proposed method's outputs. The listeners had to choose on a scale of 1-No difference to 5-Totally different the degree of similarity between the samples. The overall MOS score, 2.5, showed no significant differences between the baseline system and the proposed method, meaning that the estimation method using even a minimum amount of purely textual data is efficient and correct.

5.5 Summary

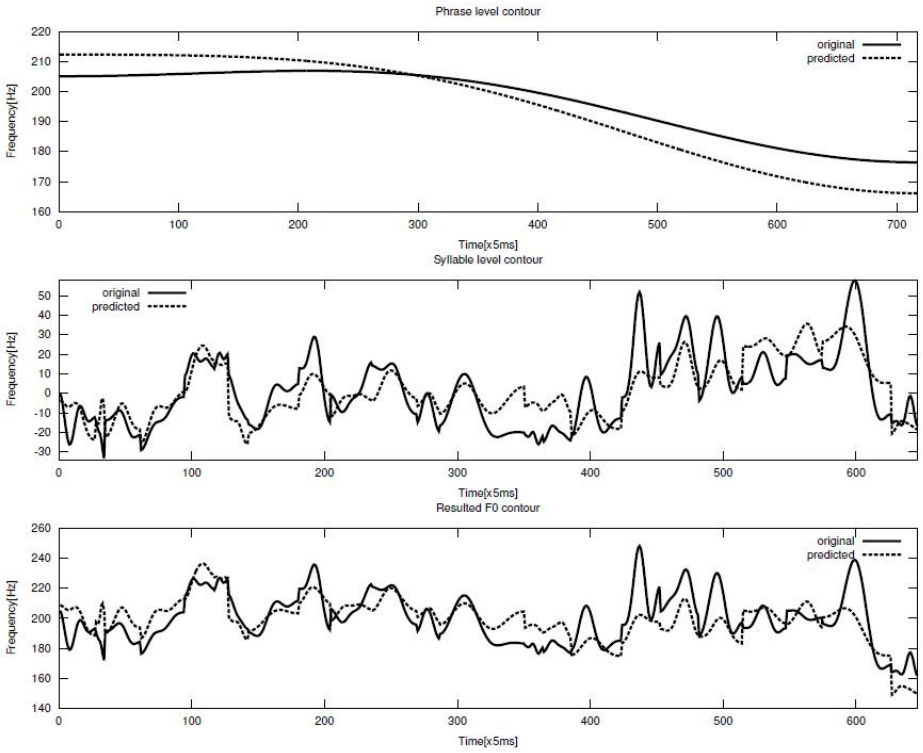
This chapter introduced a new model for F0 parametrisation using the superpositional principle and the discrete cosine transform. The problems of F0 modelling and possible solutions were first identified. Some of the most important pitch modelling and parametrising techniques were briefly presented, such as the ToBI and Fujisaki models. Because these methods require a subjective annotation and lack some essential statistical characteristics, a method based on the DCT was adopted. DCT has been shown to adhere to the superpositional model and can parametrise F0 contours with minimum error even with a reduced number of coefficients.

The proposed method uses the phrase and syllable levels of intonation, where the syllable level is the result of the original F0 contour minus the inverse DCT of the phrase level coefficients. The prediction of the DCT coefficients from the purely textual information contained in the HTS labels is then performed. Three CART trees were compared: M5

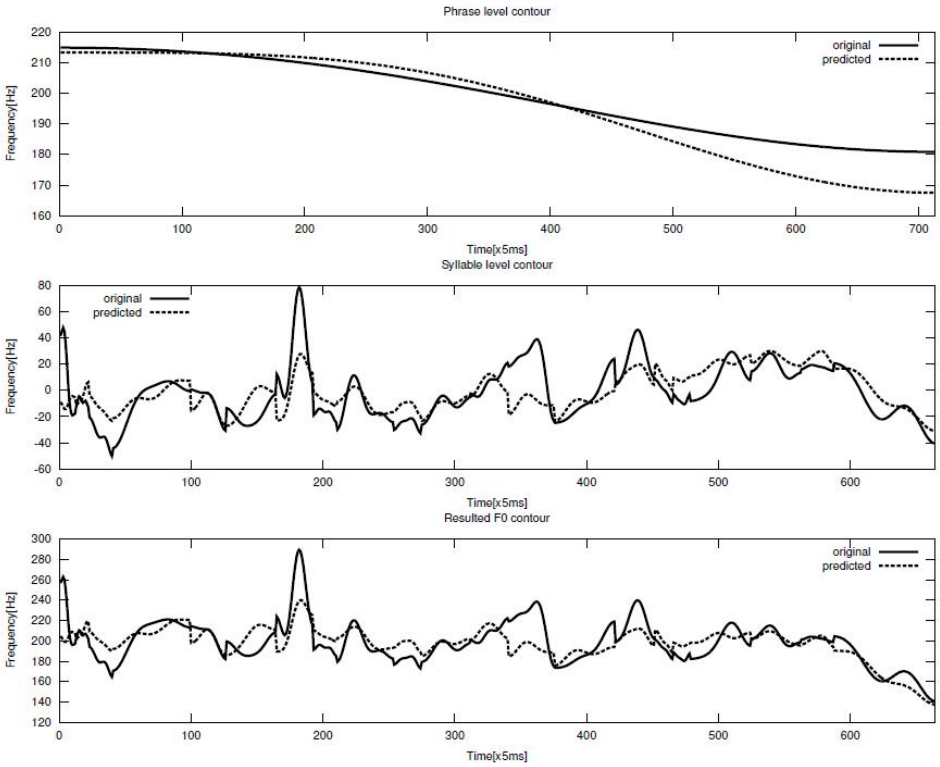
rules, additive regression and linear regression. Also, in an attempt to reduce the number of features used, a greedy stepwise attribute selection method was applied, evidentiating the best correlates between each DCT coefficient and the features used. The results of the CART training and prediction for each DCT coefficient is presented in terms of correlation and error. The best performances in a 5-fold cross validation of the training set, was obtained by the additive regression algorithm.

The algorithm was then used to predict the F0 contours of a test set of 10 utterances. The mean error obtained in the prediction stage is around 15 Hz, which is comparable to other statistical methods of F0 modelling. A listening test was also performed, and showed no significant perceptual differences between the original and estimated F0 contours.

The entire chapter is also a preliminary analysis of the DCT parametrisation capabilities in order to apply it to the interactive intonation optimisation presented in the next chapter.



(a)



(b)

Figure 5.8: Original and predicted F0 contours - utterances: (a) *Băimăreanul urăşte lipsa de punctualitate şi făţărnicia.* and (b) *În acest cămin au prioritate studenţii în ani terminali.*

Chapter 6

Optimising the F0 Contour with Interactive Non-Expert Feedback

6.1 Introduction

Over the last decade text-to-speech systems have evolved to a point where in certain scenarios, non-expert listeners cannot distinguish between human and synthetic voices with 100% accuracy. One problem still arises when trying to obtain a natural, more expressive sounding voice. Due to the subjectivity of expression and emotion realisation in speech, humans cannot objectively determine if one system is more expressive than the other.

In order to achieve a more dynamic prosody, several methods have been applied ([Tao et al., 2006], [Yamagishi et al., 2005]), some of which have had more success than others and all of which include intonation modelling as one of the key aspects. Intonation modelling refers to the manipulation of the pitch or fundamental frequency. The expressivity of speech is usually attributed to a dynamic range of pitch values. But in the design of any speech synthesis system (both concatenative and parametric), one important requirement is the flat intonation of the speech corpus, leaving limited options for the synthesised pitch contours.Therefore a solution which can extend the pitch range starting from a flat intonation input is needed.

Also, another problem is that the TTS systems have is that their output cannot be modified by a non-expert user in a simple manner. The end-user should be able to adapt

the synthetic speech to its preference with minimum feedback. The modifications should be made possible by simply comparing speech samples, as the use of other parameters involves a degree of learning and understanding on behalf of the user. The later method has been applied in some speech synthesisers, such as [Ferencz, 1997], where the user can adjust the pitch and duration of certain voice segments by inputing values for them. High-quality commercial systems can use tags (for example XML) to adjust the same parameters, or insert additional textual data[1].

An interactive intonation optimisation method based on the pitch contour parametrisation and evolution strategies is presented. The Discrete Cosine Transform (DCT) is applied to the phrase level pitch contour. Then, the genome is encoded as a vector that contains 7 most significant DCT coefficients. Based on this initial individual, new speech samples are obtained using an interactive Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm. A series of parameters involved in the process are evaluated, such as the initial standard deviation, population size, the dynamic expansion of the pitch over the generations and the naturalness and expressivity of the resulted individuals. The results provide the guidelines for the setup of an interactive optimisation system in which the users can subjectively select the individual which best suits their expectations with minimum amount of fatigue.

6.1.1 Problem statement

In this subsection some aspects of the current state-of-the-art speech synthesisers which limit the expressiveness of the result are emphasised:

Issue #1: Some of the best TTS systems benefit from the prior acquisition of a large speech corpus and in some cases extensive hand labelling and rule-based intonation. But this implies a large amount of effort and resources, which are not available for the majority of languages.

Issue #2: Most of the current TTS systems provide the user with a single unchangeable result which can sometimes lack the emphasis or expressivity the user might have hoped for.

[1]Loquendo `http://www.loquendo.com/en/demo-center/interactive-tts-demo/` allows the user to insert tags like *item=Cry_01* for specific intonational phrases or emotions

Issue #3: If the results of a system can be improved, it usually implies either additional annotation of the text or a trained specialist required to rebuild most or all of the synthesis system.

Issue #4: Lately, there have been studies concerning more objective evaluations of the speech synthesis, but in the end the human is the one to evaluate the result and this is done in a purely subjective manner.

6.2 Evolutionary Algorithms and Strategies

Evolutionary computation is a subclass of artificial intelligence, or more specifically of the computational intelligence and addresses problems of combinational optimisation [Jong, 2006]. Evolutionary computation is based on the natural process of individual selection by using the phenomenons of gene mutation and crossover. So that the evolutionary techniques solve metaheuristic optimisation problems. Subclasses of the evolutionary computation are: evolutionary algorithms, genetic algorithms, evolution strategies, genetic programing, swarm intelligence, ant-colony optimisation and swarm optimisation of the particles.

Evolutionary algorithms (EA) form a subclass of the evolutionary computation and operate with a population of potential individuals applying the survival of the fittest principle to produce better approximations of the solution. In each generation a new set of approximations is created by using a *fitness* function. The new population is then combined using genetic operators [Geatbox-webpage, 2011]. These operators model the natural processes of selection, recombination, mutation, migration, localisation and vicinity.

The problems solved by this type of algorithms refer to search and optimisation methods, problems that can be described using an objective fitness function. The possible solutions of an optimisation problem are the population's individuals, and the cost function determines the environment the individuals live in. In this process, the mutations and recombinations ofer the necessary diversity, and the selection imposes a rise in the quality of the individuals. Evolutionary algorithms differ from the traditional optimisation methods through the following [Rutkowski, 2008]:

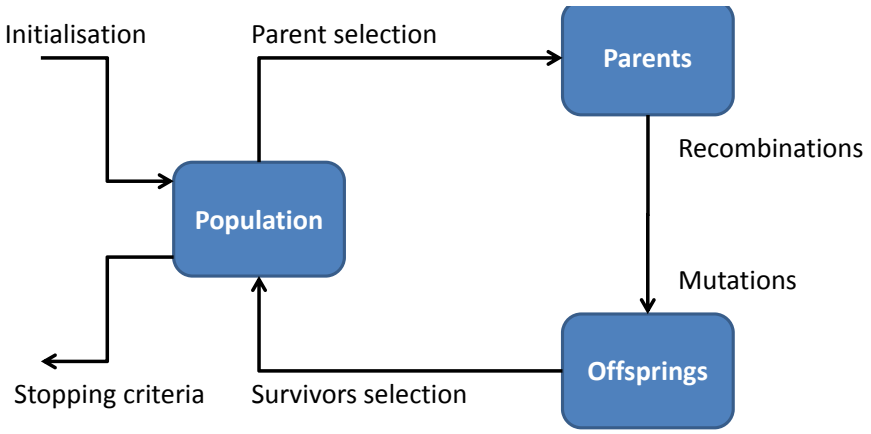


Figure 6.1: Block diagram of an evolutionary algorithm.

1. EA do not process the task parameters directly, but their coded form
2. EA realise the search starting from a population of points and not from a individual point
3. EA use just objective functions and not their derivatives or other additional information
4. EA use probabilistic selection rules at the expense of deterministic rules

The **concepts** with which the evolutionary algorithms operate are the same as in genetics [Rutkowski, 2008]:

Population - a set of fixed dimension individuals.

Individuals - of a population in a genetic algorithm are tasks coded as chromosomes, which mean solutions or points in the search space. The individuals are sometimes called *organisms*.

Chromosomes - or *chains* or *code sequences* - are ordered sequences of genes.

Gene - or *characteristic, sign, detector* - represents a single element of the genotype, of the chromosomes in particular.

Genotype - or *structure* - is a set of chromosomes of a particular individual. So that the individuals of a population can be genotypes or a single chromosome (in the

case that the genotype is composed of a single chromosome, case oftenly used n implementation)

Fenotype - is a set of values corresponding to a certain genotype, which is a decoded structure and represents a set of task parameters (a solution or a point the search space).

Allel - the value of a gene, also called the characteristic value or the characteristic variant.

Locus - is an indicator of the position of a gene in the chain, or the chromosome.

The **components** of an evolutionary algorithm [Eiben and Smith, 2010] are described below and also presented in Fig. 6.1:

Representation - specifying a link between fenotype and genotype. The elements of the fenotype are called *candidate solutions*, and the individuals define points in the search space. The search space is also called fenotype space. In the genotype, the elements are called *chromosomes*, and the determined space is a genotype space.

Fitness function - has the role to represent the adaptation requisites. It forms a selection base and thus allows improvements. In other words, it determines what an improvement means. Technically, it is a procedure or a function which attributes a measure of the genotype quality. It is called adaptation or evaluation function. It is a measure of adaptability of an individual within the population. The function is extremely important because based on its value the best adapted individuals are selected, meaning the individuals with the highest fitness, according with the principle of the *survival of the fittest* ir *survival of the strongest*. It has a mojor impact on the results of the algorithm and has to be optimally defined. In optimisation problems, the fitness function is called *objective function*. In minimisation problems, the objective function is transformed into a maximisation function. In evolutionary algorithms, at each iteration, the fitness of the population's individuals is determined by the value of the fitness function, and using this value, the new generation is created. This population is composed of a set of probable solutions to the problem. An iteration of the algorithm is called *generation*, and the generated population is called *the new generation* or *offspring population*.

Population - has the role to maintain the representations of the possible solutions. A population is a multiset of the genotype. The populations represents the evolution unit. The individuals are static objects that do not modify or adapt, the population does this. In more sofisticated EA, the population has an additional spatial structure, by defining a distance or a vicinity relation. In most of the cases the population dimension is constant. *The diversity* of a population is a measure of the number of different solutions, that can be determined by the distinct values of the fitness function or by the number of different fenotypes or genotypes.

Parent selection mechanism - distinguishes between individuals based on their quality, to allow for the best individuals to become parents of the new generation. An individual is a *parent* if it has been selected in order to realise a variation and create an offspring.Along with the survivor selection, parent selection is responsible for the improvement of the quality of the system. In evolutionary computation, the selection is usually probabilistic, so that the individuals with a higher quality have a higher chance of becoming parents.

Variation, recombination and mutation operators - have the role of creating new individuals. It is similar to the generation of new candidate solutions. The variation operators are split in two categories: *mutations* and *recombinations*. **The mutations** generate random impartial modifications. They are stochastic operators – the offspring is the results of a series of random choices. **Recombination** – determines the generation of offspring using the recombination or crossover of the parents' genotype. The choice of the parents' genotype to be combined is random.

Survivor selection mechanism - determines a hierarchy of the individuals based on their quality. It is also called *replacement*. The decision is made based on the values of the fitness functions. As opposed to the parent selection, the survivor selection is often deterministic, the individuals are ordered, and the first N are selected to generate the next generation.

Initialisation - is usually done at random. Some conditions can be applied to determine an initial population with a good fitness.

Stopping criteria - depends on the chosen method: if there is a certain optimal value of the fitness, then reaching this level determines the algorithm to stop. If this level cannot be achieved, other conditions can be applied, such as the number of iterations or generations, or when the diversity of the population is below a threshold.

The most important evolutionary algorithms are: *genetic algorithms, evolutionary computation* and *evolution strategies*. These types of algorithms will be briefly presented in the following sections, with an emphasis on evolution strategies which are applied in the proposed method.

6.2.1 Genetic Algorithms

Genetic algorithms (GA) represent a heuristic search method which imitates the processes of natural selection. They are used in the search and optimisation solutions. GA follow closely the steps of the evolutionary algorithms: intialisation, selection, reproduction and stopping criteria[Jong, 2006, Rutkowski, 2008]. The detailed steps are as follows:

- random generation of m parents
- fitness function computation for each of the population's individuals
- the definition of the selection probabilities for each of the parents so that they are directly proportional with the fitness function
- the generation of m offspring by the probabilistic selection of the parents
- selection of just the offspring, setting aside the parents

It is important to note the proportionality of the fitness function. Using its value, the medium fitness individuals will produce on average a single offspring, and the ones with an above average fitness will produce more than an offspring. On the other hand, because the parents are not present in the next generation, it is possible that the mean value of the fitness to drop, which is undesirable in complex problems.

6.2.2 Evolutionary Computation

Initially, evolutionary programing (EP) was developed in the context of determining the grammar of unknown languages. The grammar was modelled through the use of a fi-

nite state machine undergoing evolution. The results were promising, but evolutionary programming became more useful when it introduces numeric optimisation.

This method is similar to evolution strategies. In evolutionary programming algorithms, the new population is create using mutation of each parent individual. In an evolution strategy on the other hand, each individual has the same chance of appearance in a temporary population unto which genetic operators are applied. The new parent population is created using a classification selection which is applied both to old populations' individuals and to the new generation. Individuals' mutations in EP represents a random perturbation of the value of a gene [Rutkowski, 2008]. So that the main variation method is mutation, the population members are considered as being part of a certain species, but not the same, so that each parent creates an offspring. EP implements an elitist strategy of survival, in which only the top 50% of the individuals survive, which determines an increase in the fitness by reducing diversity, but in the same time can lead to suboptimal solutions [Jong, 2006].

6.2.3 Evolution Strategies

Evolution strategies (ES) are an optimisation technique based on the ideas of evolution and adaptation. They use problem dependent representation, and the main search operators are mutation and selection [Jong, 2006].

In the same way as the genetic algorithms, ES operate over the potential solution population and use the selection principle and survival of the fittest. The differences rely mainly in the individual representation method. Evolution strategies use real number vectors, while the genetic algorithms use binary valued vectors. On the other hand, genetic algorithms select an offspring population equal to the parent population. The probability of selecting an individual depends on the value of the fitness function. The result can contain even the weakest of the chromosomes. Within ES, a temporary population is created in order to select the best individuals, and this process is not recurrent. The selection is deterministic.

The third difference lies in the fact that in evolution strategies the order of the selection and recombination processes is reversed. Recombination is the first process, and then the selection. An offspring is the result of a crossover between two parents and some

mutations. GA selects first the individuals and only after that applies the genetic operators. A final difference is that the parameters of the genetic algorithms remain constant from one generation to the other, while in ES, these parameters are updated constantly [Rutkowski, 2008].

ES have a few basic types:

$(1+1)$ **strategy** - a single basis chromosome is processed. The algorithm starts by randomly selecting the values for a vector X's components. In each generation a new Y individual is created as a result of a mutation. The fitness of the individuals is compared, and the highest value becomes the basic chromosome of the next generation. There is no crossover.

$(\mu+\lambda)$ **strategy** - the algorithm starts from a random parent generation which contains μ individuals. A temporary generation T is created through reproduction. Reproduction is the random selection of λ individuals from the initial population and placing them in the temporary population. The individuals of the T generation suffer crossover and mutation operations, resulting a new offspring population O with a dimension λ. The best μ individuals are selected from $P \cup O$ and a new parent generation of dimension μ is created.

(μ,λ) **strategy** - the difference between this strategy and the previous one is that the new P generation with μ individuals is selected just out of the first λ individuals of the O generation. The condition $\mu > \lambda$ has to be satisfied. The advantage over the $(\mu+\lambda)$ strategy is that in the previous one, the popualtion could be dominated by a single high fitness individual, but with too high or too low standard deviation. In (μ,λ) the old individuals are not transfered to the new reproduction base.

CMA-ES

The CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) is a stochastic method for real-parameter (continuous domain) optimization of non-linear, non-convex functions. CMA-ES was proposed by Hansen and Ostermeier [Hansen and Ostermeier, 1996] as an evolutionary algorithm to solve unconstrained or bounded constraint, non-linear optimisation problems defined in a continuous domain. In an evolutionary algorithm, a *population*

of genetic representations of the solution space, called *individuals*, is updated over a series of iterations, called *generations*. At each generation, the best individuals are selected as *parents* for the next generation. The function used to evaluate individuals is called the *fitness* function.

The search space is explored according to the genetic operations used to update the individuals in the parent population and generate new offspring. In the case of evolution strategy (ES), the selection and mutation operators are primarily used, in contrast to the genetic algorithm (GA) proposed by Holland [Holland, 1975], which considers a third operator – crossover. Also, in GA the number of mutated genes per individual is determined by the *mutation probability*, while in ES mutation is applied to all genes, slightly and at random.

If mutation is according to a multivariate normal distribution of mean m and covariance matrix C, then CMA-ES is a method to estimate C in order to minimise the search cost (number of evaluations). First, for the mean vector $m \in \mathbb{R}^n$, which is assimilated to the preferred solution, new individuals are sampled according to the normal distribution described by $C \in \mathbb{R}^{n \times n}$:

$$x_i = m + \sigma y_i \tag{6.1}$$

$$y_i \sim N_i(0, C), i = 1..\lambda$$

where λ is the size of the offspring population and $\sigma \in \mathbb{R}_+$ is the step size.

Second, sampled individuals are evaluated using the defined fitness function and the new population is selected. There are two widely used strategies for selection: $(\mu + \lambda)$-ES and (μ, λ)-ES, where μ represents the size of the parent population. In $(\mu + \lambda)$-ES, to keep the population constant, the λ worst individuals are discarded after the sampling process. In (μ, λ)-ES all the parent individuals are discarded from the new population in favour of the λ new offspring.

Third, m, C and σ are updated. In the case of (μ, λ)-ES, which is the strategy chosen for implementation in this solution, the new mean is calculated as follows:

$$m = \sum_{i=1}^{\mu} w_i x_i \tag{6.2}$$

$$w_1 \geq .. \geq w_\mu, \sum_{i=1}^{\mu} w_i = 1$$

where x_i is the i-th ranked solution vector ($f(x_1) \leq .. \leq f(x_\lambda)$) and w_i is the weight for sample x_i.

The covariance matrix C determines the shape of the distribution ellipsoid and it is updated to increase the likelihood of previously successful steps. Details about updating C and σ can be found in [Hansen, 2005] .

6.3 Interactive Intonation Optimisation

The issues presented in the Problem statement section at the beginning of this chapter are partially addressed through the method presented next [Stan et al., 2011a]. The method is as follows: given the output of a synthesiser, the user can opt for a further enhancement of its intonation. The system then evaluates the initial pitch contour and outputs a small number of different versions of the same utterance. Provided the user subjectively selects the best individual in each set, the next generation is built starting from this selection. The *dialogue* stops when the user considers one of a generation's individual satisfactory. The solution for the pitch parametrisation is the Discrete Cosine Transform (DCT) and for the interactive step, the Covariance Matrix Adaptation-Evolution Strategy (CMA-ES).

This method is useful in the situation where non-expert users would like to change the output of a speech synthesiser to their preference. Also, under resourced languages or limited availability of speech corpora could benefit from such a method. The prosodic enhancements selected by the user could provide long-term feedback for the developer or could lead to a *user-adaptive* speech synthesis system.

6.3.1 Related Work

To the best of the author's knowledge, evolution strategies have not been previously applied to speech synthesis. However, the related genetic algorithms have been used in articulatory [D'Este and Bakker, 2010] or neural networks based [Moisa et al., 2001] speech synthesisers. [Moisa et al., 2001] presents a method of supervised training of neural networks using evolutionary algorithms. The network structure is a hyper-sphere.

A study of interactive genetic algorithms applied to emotional speech synthesis is presented in [Lv et al., 2009]. The authors use the XML annotation of prosody in Microsoft Speech SDK and try to convert neutral speech to one of the six basic emotions: *happiness, anger, fear, disgust, surprise* and *sadness*. The XML tags of the synthesised speech comprise the genome. Listeners are asked to select among 10 speech samples at each generation and to stop when they consider the emotion in one of the speech samples consistent with the desired one. The results are then compared with an expert emotional speech synthesis system. Listeners had to rate on average 100 versions of the speech sample, going as far as the 10th generation of the algorithm. The prosody adaptation is realised at word level and not at phrase or utterance level.

Some other applications of the evolution strategies can be found in the conversion of the quality of the synthesised voices [Sato, 2005], or in the optimisation of the Markov models for signal modelling [Huda et al., 2009]. Interactive evolutionary computation has, on the other hand, been applied to music synthesis [McDermott et al., 2010], and music composition [Fukumoto, 2010], [Marques et al., 2010].

6.3.2 DCT Parametrisation of the phrase level F0 Contour

The method proposed addresses the issue of modelling the **phrase level** intonation, or trend. Starting from a flat intonation, a more dynamic and expressive contour is derived. Therefore, it is considered that the phrase layer is represented by the inverse DCT transform of the DCT1 to DCT7 coefficients of the pitch DCT. This assumption is also supported by the results presented in [Teutenberg et al., 2008] and previous chapter. DCT0 represents the mean of the curve and in this case it is speaker dependent. Using DCT0 in the genome encoding would undesirably change the pitch of the speaker, the focus being on the overall trend of the phrase intonation. The phrase level is then subtracted from the overall contour, and the result is retained and will be referred to as *high level pitch information*. Fig. 6.2 presents an example of a pitch contour, the phrase level contour based on the inverse DCT of the DCT1-DCT7 coefficients and the high level pitch information. It can be observed that the phrase level contour represents the relative trend of the voiced segments intonation, while the high level information has a relatively flat contour with variations given by the word, syllable and phoneme levels.

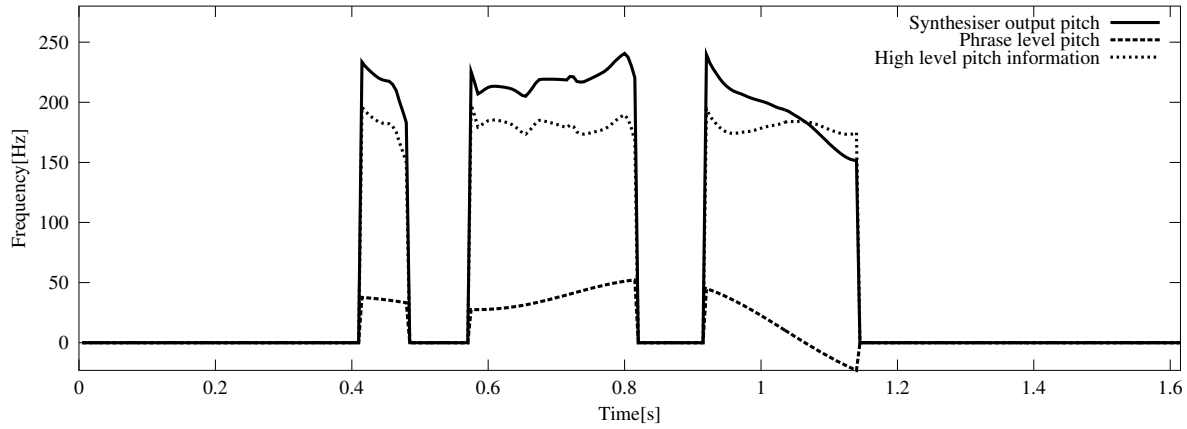


Figure 6.2: An example of a pitch contour decomposition into phrase level and high level pitch information. The phrase level contour is based on the inverse DCT of DCT1-DCT7 coefficients – utterance "Ce mai faci?" ("How are you?").

Because DCT cannot parametrise fast variations with a small number of coefficients, the unvoiced segments of the F0 contour were interpolated using a cubic function. During the interactive step, the inverse DCT transform is applied to the winner's genome, the high level pitch information is added and the speech is synthesised using the resulted F0 contour.

6.3.3 Proposed solution

Combining the potential of the DCT parametrisation and evolution strategies, an interactive solution for the intonation optimisation problem is introduced, and requires no previous specific knowledge of speech technology. To achieve this, three issues need to be solved: *1)* generate relevant synthetic speech samples for a user to chose from, *2)* minimise user fatigue and *3)* apply the user feedback to improve the intonation of the utterance.

The first issue is solved by using CMA-ES to generate different speech samples, normally distributed around the baseline output of a Romanian speech synthesis system [Stan et al., 2011b] based on HTS (Hidden Markov Models Speech Synthesis System) [Zen et al., 2007a]. The *genome* is encoded using a vector of 7 genes, where each gene stores the value of a DCT coefficient, from DCT1 to DCT7. It starts with an initial mean vector m that stores the DCT coefficients of the F0 phrase level generated by the

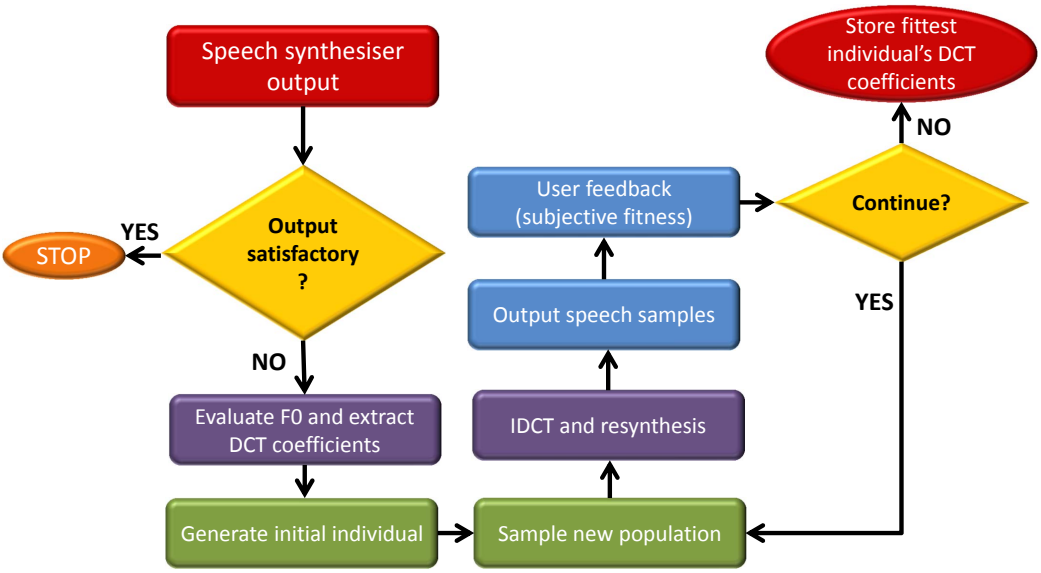


Figure 6.3: Flow chart of the proposed method for the interactive intonation optimisation algorithm.

HTS system and an initial covariance matrix $C = I \in \mathbb{R}^{7\times7}$. In each generation, new individuals are sampled according to Eq. (6.1).

In the next step, the user needs to evaluate generated individuals. If the population size is too large, the user may get tired before a suitable individual is found or might not spot significant differences between the individuals. On the other hand, if the population size is too small and the search space is not properly explored, a suitable individual may not be found. CMA-ES is known to converge faster even with smaller population than other evolutionary algorithms, but it was not previously applied to solve interactive problems. On the other hand, interactive genetic algorithms (IGA) have been extensively studied, but do not converge as fast as CMA-ES for non-linear non-convex problems. Faster convergence means fewer evaluations, therefore reducing user fatigue.

For the interactive version of CMA-ES, a *single elimination tournament* fitness function [Panait and Luke, 2002] was used. In this case, the individuals are paired at random and play one game per pair. Losers of the game are eliminated from the tournament. The process repeats until a single champion is left. The fitness value of each individual is equal to the number of played games. Each pair of individuals is presented to the user in the form of two speech samples. Being a subjective evaluation, the choice would best

suit the user's requirements, thus giving the winner of a population.

The fitness value is used by CMA-ES to update mean vector m, the covariance matrix C and the standard deviation σ. A new population of individuals is sampled based on the updated values and the process repeats. The flow chart of the proposed method is presented in Fig. 6.3.

6.4 Evaluation

The results presented below focus on establishing the correct scenario for the interactive application and on the ease of use on behalf of the listeners/users. This implies the evaluation of several parameters involved, such as: *initial standard deviation of the population* – gives the amount of dynamic expansion of pitch –, *the population size* – determines the number of samples the user has to evaluate in each generation, *the expressivity and naturalness of the generated individuals* – assures correct values for the pitch contour.

6.4.1 Experiment 1 - Initial Standard Deviation of the Population

As a preliminary step in defining the standard deviation of the population, some of the statistical result presented in chapter 5 are reviewed. These include the analysis all the DCT coefficients within the *rnd1* subset of the Romanian Speech Synthesis corpus [Stan et al., 2011b]. The number of phrases within this subset is 730 with an average length of 1.7 seconds. The intonation of the speech is flat, declarative. DCT0 was included as well for an overall view as it represents the mean of the pitch contour and it is speaker dependent. This coefficient was not used in the estimation of the phrase level contour. The means and standard deviations of the coefficients are presented in Table 6.1, also including their corresponding F0 values in Hz. From Table 6.1 it can be observed that DCT0 can be a good indicator of the speaker's fundamental frequency, with an average value of 254Hz over the entire speech corpus.

The average pitch contour resulted from the mean values of the DCT coefficients and the average duration of the *rnd1* subset is shown in Fig. 6.4. The pitch contour corresponds to a declarative intonational pattern corresponding to the speech data analysed.

Table 6.1: Means and standard deviation of the DCT coefficients in *rnd1* subset with corresponding variations in Hz for an average length of 1.7 seconds.

'Coefficient	Mean	Mean F0 [Hz]	Standard deviation	Maximum F0 deviation [Hz]	
				- 1 std dev	+1 std dev
DCT0	4690.300	251-257	1318.300	179-186	322-329
DCT1	331.750	± 4	185.850	±12	±40
DCT2	-95.087	±7	197.470	±22	±7
DCT3	168.270	±12	161.030	±0.55	±25
DCT4	-57.100	±4	151.600	±16	±7
DCT5	94.427	±7	130.150	±2	±17
DCT6	-22.312	±1	123.020	±11	±7
DCT7	67.095	±5	110.370	±3	±13

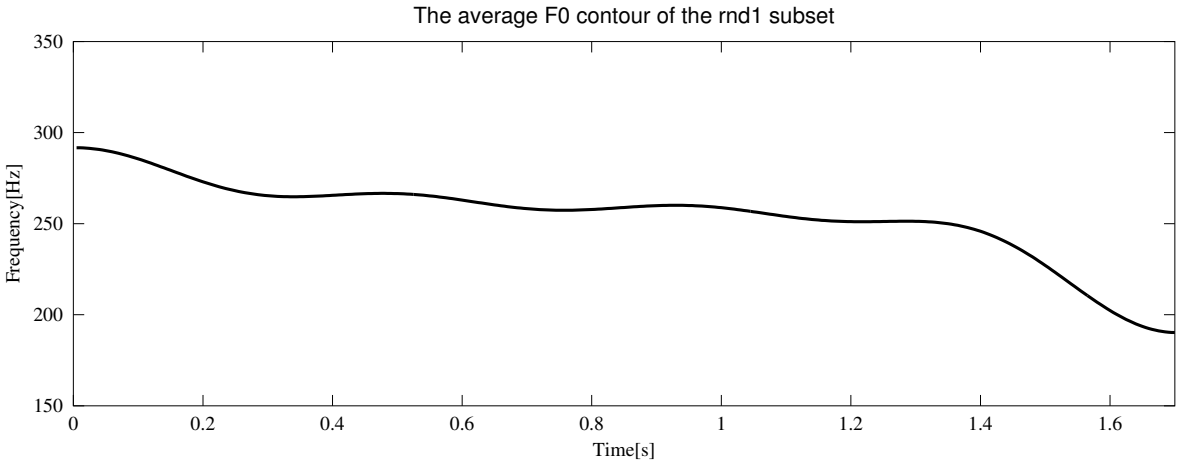


Figure 6.4: Result of the pitch contour generated from the mean values of the DCT0-DCT7 coefficients within the *rnd1* subset, and an average phrase length of 1.7 seconds

DCT1 has the most important influence in the F0 contour after DCT0. The mean value of the DCT1 coefficient is 331.75 with a standard deviation of 185.85 and the maximum F0 variation is given by the *+1 std. dev.* (i.e. 331.75+185.85 = 517.6) of around 40 Hz. One of the issues addressed in this thesis is the expansion of the pitch range. This means that having a standard deviation of the flat intonation speech corpus, a higher value for it should be imposed while generating new speech samples, but it should not go up to the point where the generated pitch contours contain F0 values which are not natural. In Fig. 6.5 the third generation for an initial standard deviation of 150 and 350 respectively is compared. It can be observed in the 350 case that individual 3 has F0 values going

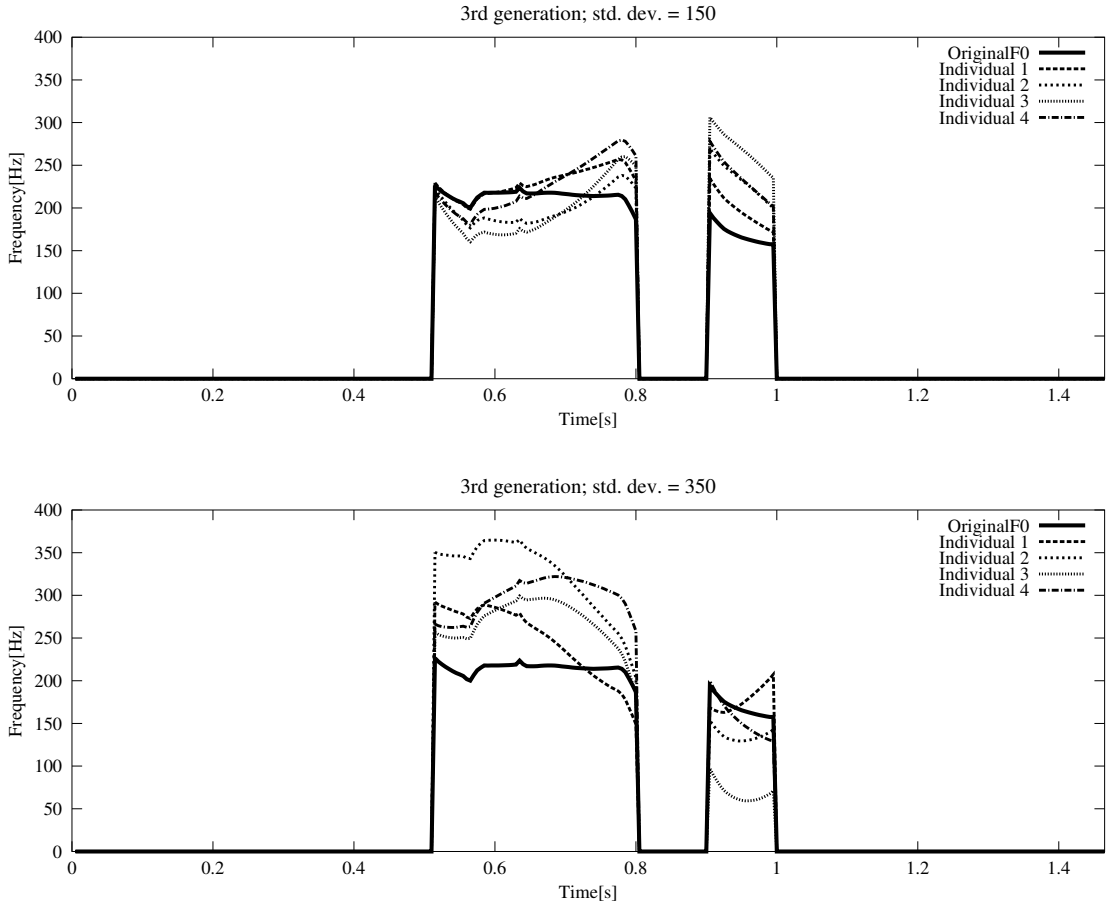


Figure 6.5: The 3rd generation population of the F0 contour, with an initial standard deviation of 150 and 350 respectively. Original F0 represents the pitch contour produced by the synthesiser – utterance "Ce mai faci?" ("How are you?")

as low as 50 Hz – unnatural, while for a standard deviation of 150, the F0 contours do not vary too much from the original one and lead to a less dynamic output. Given these results, a standard deviation of 250 was selected. An important aspect to be noticed from Table 6.1 is that all the 7 coefficients have approximately the same standard deviation. This means that imposing a variation based on DCT1 does not exceed natural values for the rest of the coefficients.

6.4.2 Experiment 2 - Population Size

The single elimination tournament fitness used to evaluate the individuals requires the user to provide feedback for $n - 1$ games, where n is the population size. So that the population size has a great importance in setting up the interactive application. Several

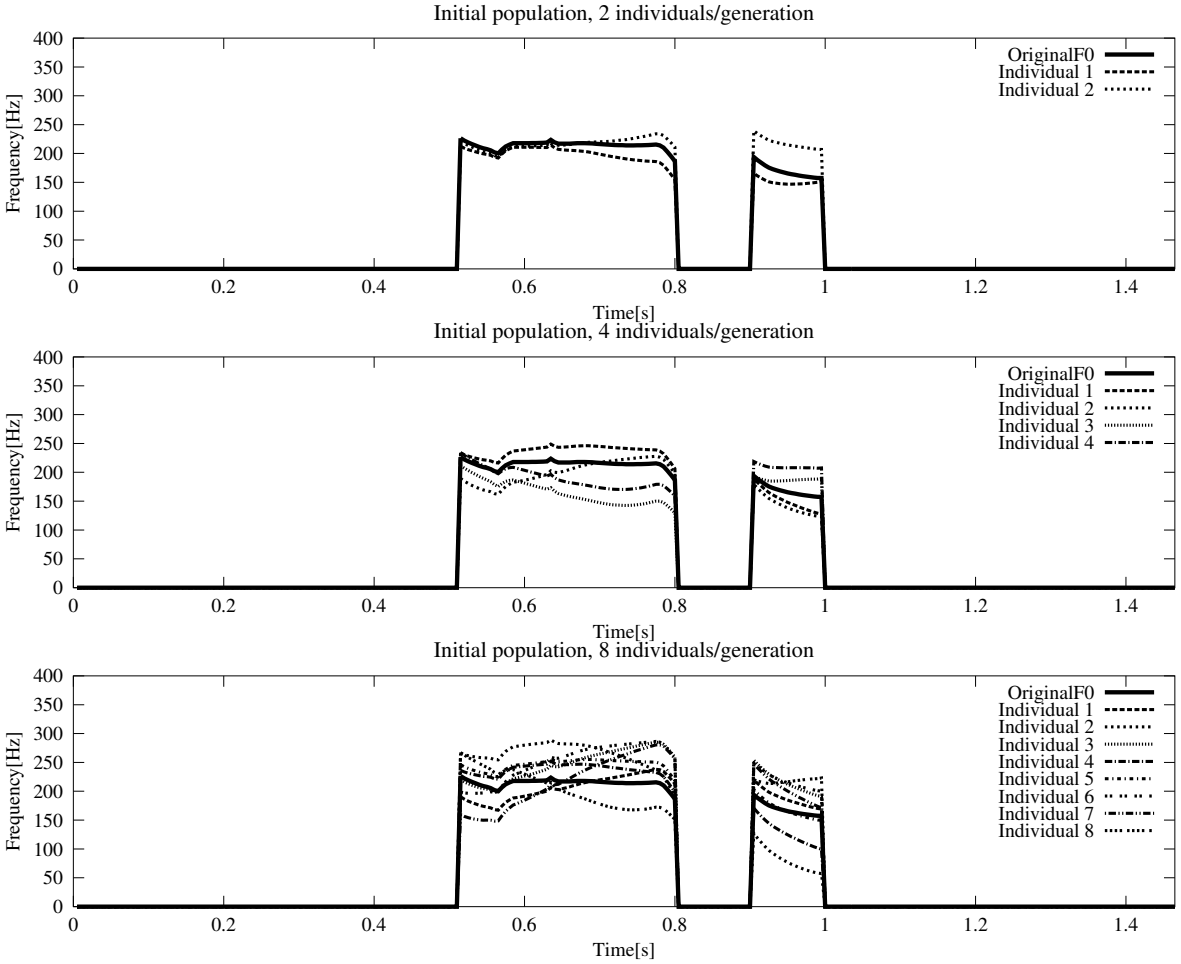


Figure 6.6: Population size variation. Original F0 represents the pitch contour produced by the synthesiser – utterance "Ce mai faci?" ("How are you?").

values have been selected for it and the results are shown in Fig. 6.6. Although the highest the number of individuals the more samples the user can choose from, this is not necessarily a good thing in the context of user fatigue. But having only 2 individuals does not offer enough options for the user to choose from. Therefore the use of 4 individuals per generation is suggested as a compromise between sample variability and user fatigue.

6.4.3 Experiment 3 - Dynamic Expansion of the Pitch

Another evaluation is the observation of the modification of the pitch contour from one generation to the other. Fig. 6.7 presents the variation of F0 from the initial population to the third. It can be observed that starting with a rather flat contour, by the third generation the dynamics of the pitch are much more expanded, resulting a higher intonation

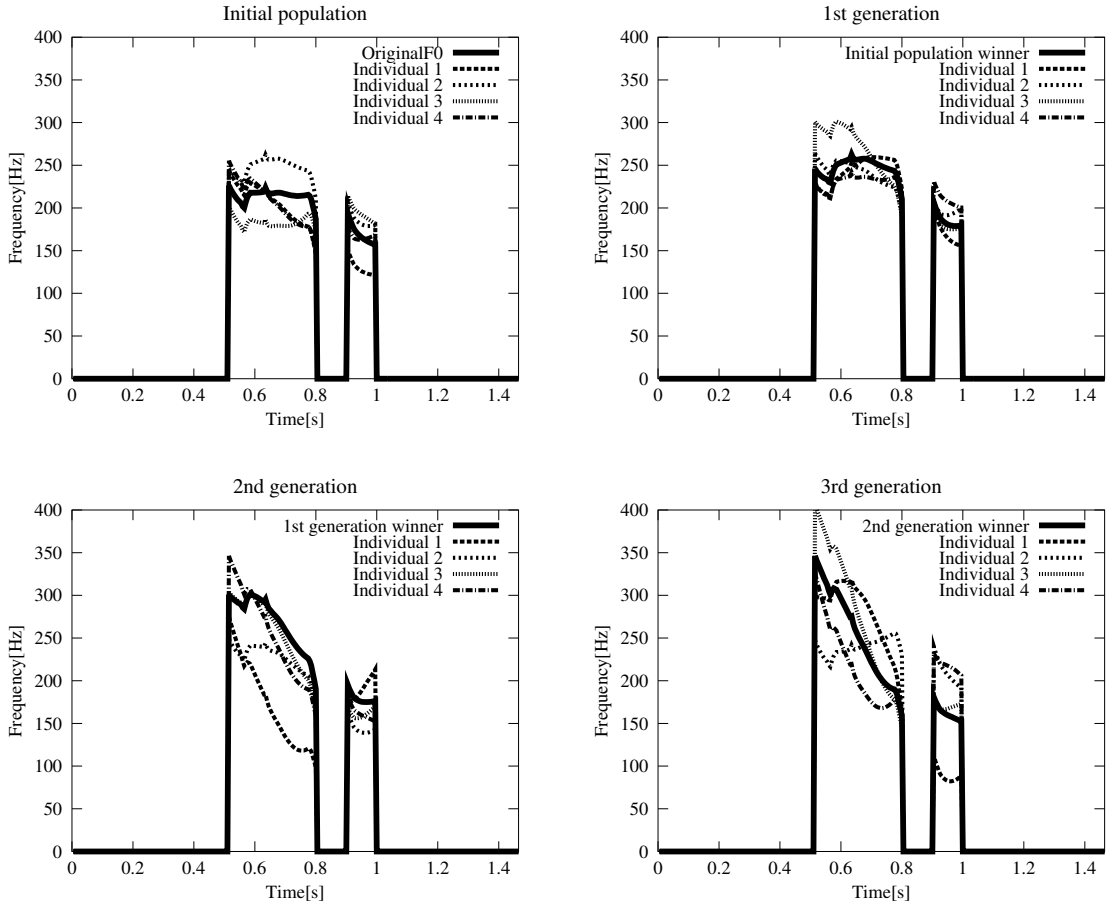


Figure 6.7: Evolution of the F0 contour over 3 generations, standard deviation = 250, phrase "Ce mai faci?" ("How are you?"). Original F0 represents the pitch contour produced by the synthesiser.

variability within and between generations. It is also interesting to observe the phrase level contours (Fig. 6.8). This is a more relevant evaluation as it shows the different trends generated by CMA-ES and the trend selected by the user in each generation. The selected trend can be used in the adaptation of the overall synthesis. In the example, the user selected an intonation with a high starting point and a descending slope, while another user could have chosen individual 1 which contains an initial ascending slope.

6.4.4 Experiment 4 - Listening Test

In order to establish the naturalness of the generated individuals and the enhanced expressivity of the winners of each generation, a small listening test was conducted. At first, a user was asked to select the winners over 4 generations for 10 phrases. Initial

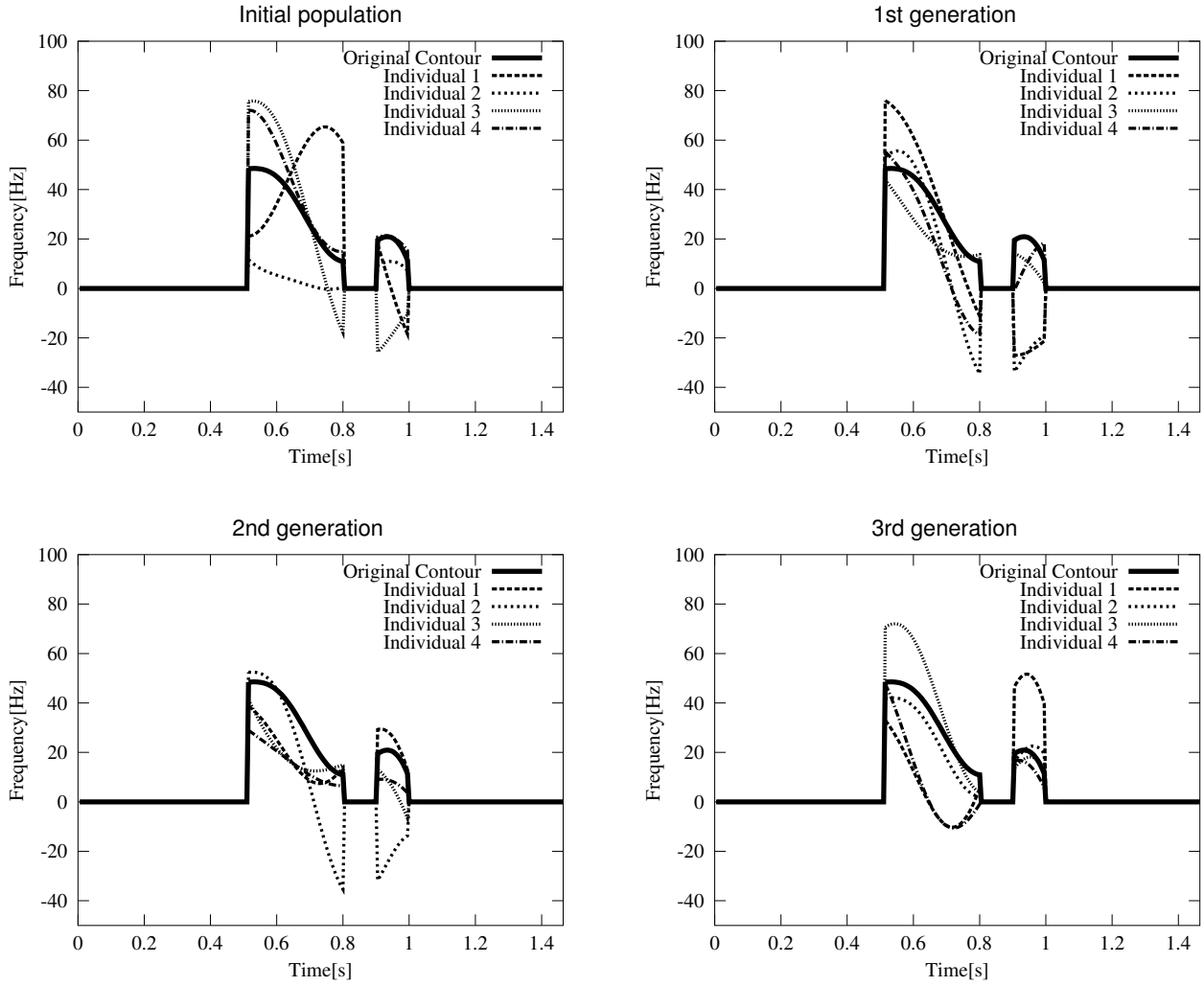


Figure 6.8: Evolution of the phrase contour trend over 3 generations for the utterance "Ce mai faci" ("How are you"). Original contour represents the pitch contour produced by the synthesiser.

standard deviation was 250 and with a population size of 4. Then 10 listeners had to attribute Mean Opinion Scores (MOS) for the samples in two categories: *Naturalness* – the generated samples were compared to original recordings on a scale of [1 - Unnatural] to [5 - Natural]. All the individuals of the four generations were presented. *Expressivity* – the winners of each generation were compared to the correspondent synthesised versions of them. The listeners had to mark on a scale of [1-Less expressive] to [5-More expressive] the generated samples in comparison to the synthesiser's output.

The results of the test are presented in Fig. 6.9. In the naturalness test, all the generations achieved a relatively high MOS score, with some minor differences for the 4^{th}

generation. The expressivity test reveals the fact that all the winning samples are more expressive than the originally synthesised one. The test preliminary conclude the advantages of this method. While maintaining the naturalness of the speech, its expressivity is enhanced. Examples of the speech samples generated by this method can be found at `http://www.romaniantts.com/nicso2011`.

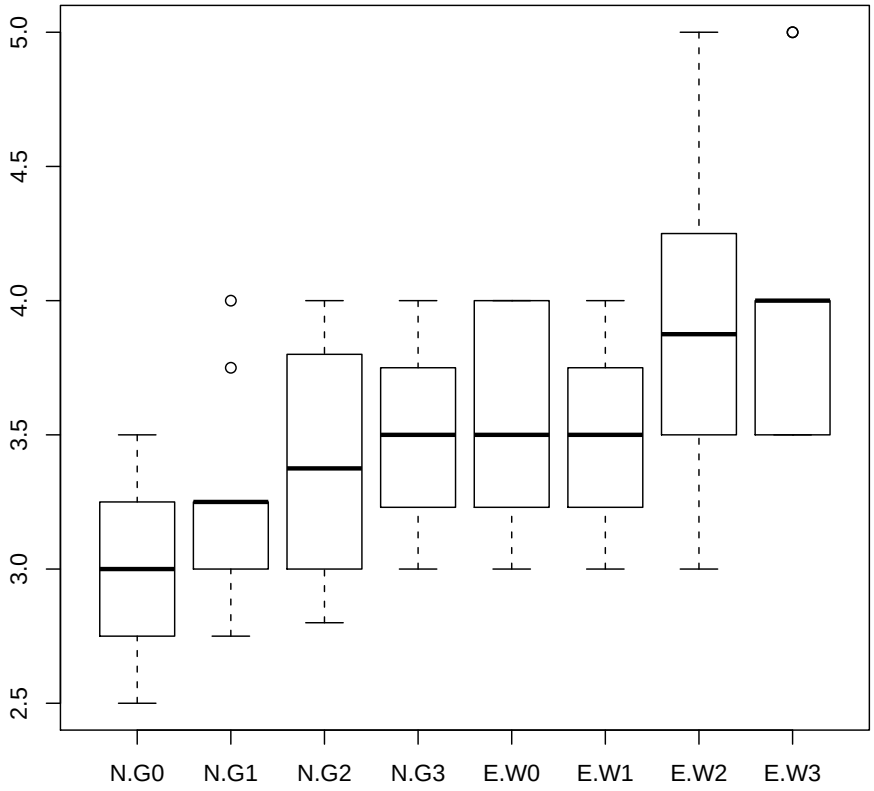


Figure 6.9: Results of the interactive intonation optimisation listening test. N-Gx represent the results for the naturalness test of each generation and E-Wx represent the results for the expressivity test of each generation's winner. The graph is a box plot, where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range.

6.5 Summary

A new method for intonation optimisation of a speech synthesis system based on CMA-ES and DCT parametrisation of the pitch contour was introduced. The interactive manner of the optimisation allows the users to select an output which best suits their expectations. The novelty of the solution consists in using no prosodic annotations of the text, no deterministic rules and no predefined speaking styles. Also, to the best of the author's knowledge, this is one of the first applications of CMA-ES for an interactive problem. The method uses the output of the Romanian HTS system presented in chapter 4 and parametrises the output's F0 contour using the DCT. Evolution strategies are applied to derive new and enhanced speech samples according to the listener's choice.

The DCT parametrisation was chosen based on the results of the previous chapter, which proved that the discrete cosine transform is capable of both modelling and predicting of the F0 contour. The proposed method uses 7 DCT coefficients DCT1 to DCT7. DCT0 is left aside because it represents the mean of the curve and can be considered as the fundamental frequency of the speaker. The influence of the mean and standard deviation of the coefficients over pitch variations is also studied. Using these results, the initial standard deviation of the population is assigned, in order to concurrently maintain the naturalness of the speech samples, while expanding the dynamics of the pitch.

Because the method is designed to be used in an interactive optimisation problem, interactive evolution strategies were selected to provide a fit population of speech samples and the correct evolution of the samples from one generation to the other. To determine the optimal population size, user fatigue versus population variation were compared. This resulted in the use of 4 individuals per generation and their evolution over 3 generations was studied in order to prove that their pitch contours evolve in a satisfactory manner.

All of the previous results were applied within a listening test, in which expressivity and naturalness of the individuals of 4 successive populations were evaluated. The listeners rated the naturalness as above average, and the individuals of higher order generations were considered to be more expressive than the initial ones.

Chapter 7

Discussion and Future Work

7.1 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 majority of the accents 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.

7.2 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 param-

eters 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 are. 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.

7.3 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.

7.4 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.

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

In chapter: 3

Published Papers: [Stan et al., 2011c], [Stan and Giurgiu, 2010], [Stan, 2010]

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

In chapter: 3

Published Papers: [Stan et al., 2011c], [Stan and Giurgiu, 2010], [Stan, 2010]

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

In chapter: 4

Published Papers: [Stan et al., 2011c], [Stan, 2010]

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

In chapter: 4

Published Papers: [Stan et al., 2011c], [Stan, 2010]

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

In chapter: 5

Published Papers: [Stan and Giurgiu, 2011], [Stan, 2011a]

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

In chapter: 6

Published Papers: [Stan et al., 2011a], [Stan, 2011b]

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 annota-

tions 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

In chapter: 6

Published Papers: [Stan et al., 2011a], [Stan, 2011b]

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. 7.1 and their interdependency as in Fig. 7.2.

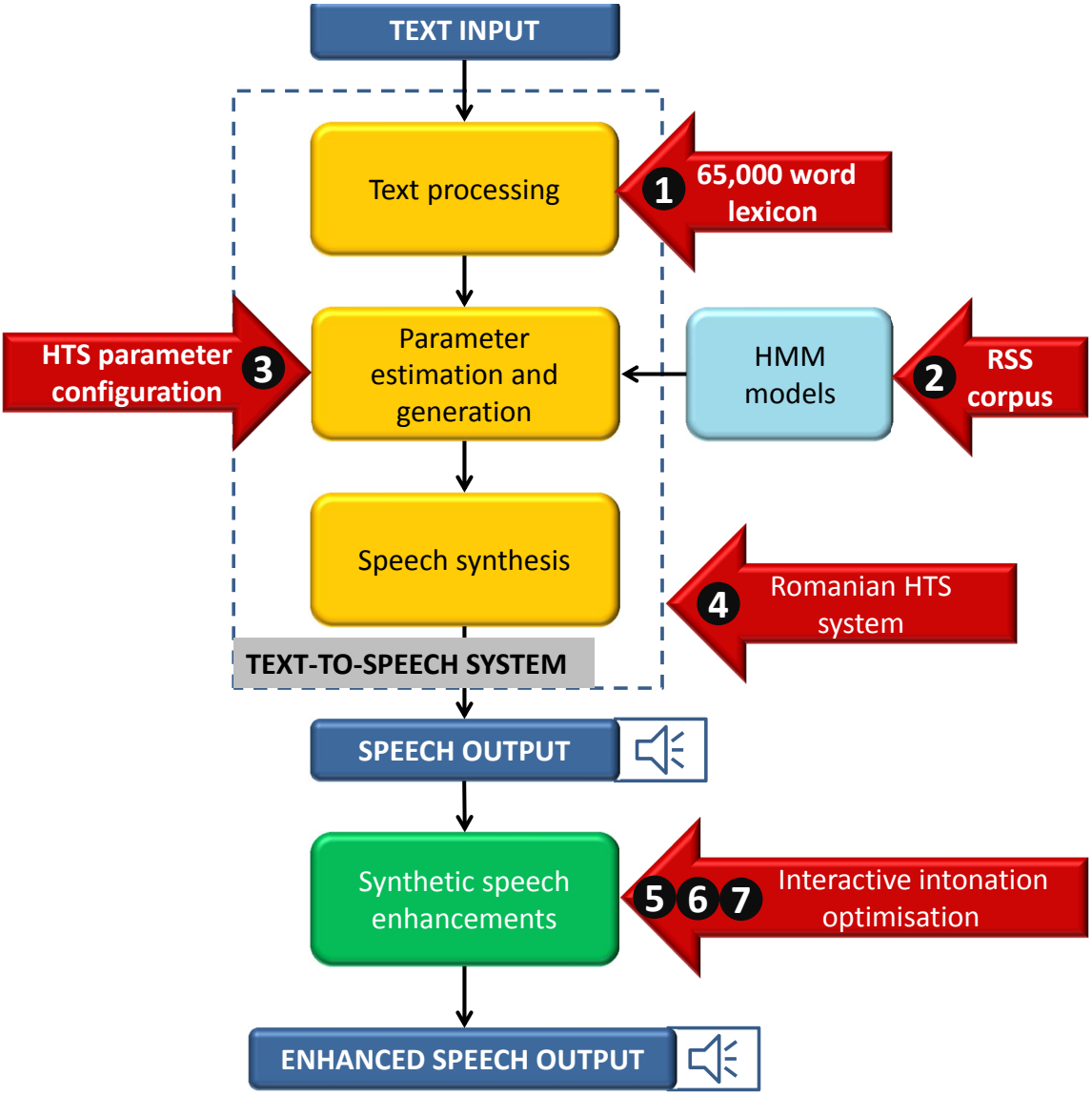


Figure 7.1: The application of the thesis contributions within the general processing scheme of an HMM-based speech synthesis system (marked with numbers from 1 to 7).

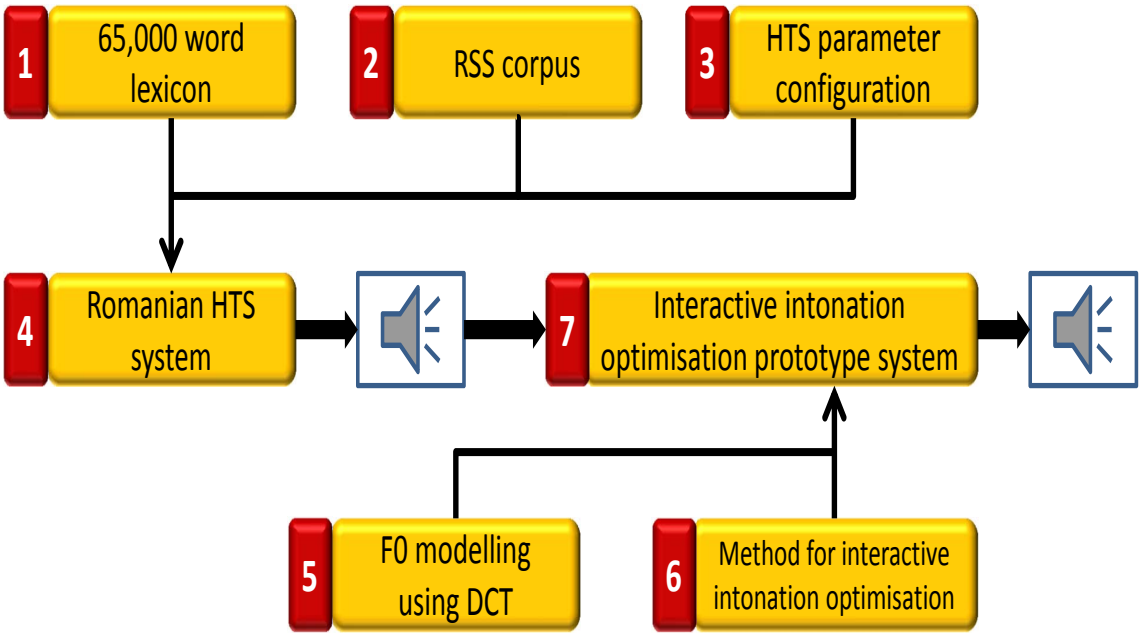


Figure 7.2: The interdependency of the thesis contributions

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 5_{th} 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 6^{th} Conference on Speech Technology and Human-Computer Dialogue, 2011

3. **Adriana STAN**, Mircea GIURGIU, *Romanian language statistics and resources for text-to-speech systems*,In Proceedings of the 9^{th} edition of the International Symposium on Electronics and Telecommunications, pg. 381-384, 2010.

4. **Adriana STAN**, *Linear Interpolation of Spectrotemporal Excitation Pattern Representations for Automatic Speech Recognition in the Presence of Noise*, In Proceedings of the 5^{th} Conference on Speech Technology and Human-Computer Dialogue, pg. 199-206, 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

Bibliography

[Acero, 1999] Acero, A. (1999). Formant analysis and synthesis using hidden Markov models. In *Proc. of Eurospeech*.

[Allen et al., 1987] Allen, J., Hunnicut, S., and Klatt, D. (1987). *From Text to Speech: the MITalk System*. Cambridge University Press.

[Apopei and Jitcă 2005] Apopei, V. and Jitcă D. (2005). Romanian Intonational Annotation Based on Tone Sequence Mode. In *Proceedings of SASM 2005*.

[Apopei and Jitcă 2006] Apopei, V. and Jitcă D. (2006). A set of Intonational Category for Romanian Speech and Text Annotation. In *Proceedings of ECIT 2006*.

[Apopei and Jitcă 2007] Apopei, V. and Jitcă D. (2007). Module for F0 Contour Generation Using as Input a Text Structured by Prosodic Information. In *Proceedings of SPED 2007*.

[Apopei and Jitcă 2008] Apopei, V. and Jitcă D. (2008). Intonational Variations for Romanian Yes-No Questions. In *Proceedings of the 5th European Conference on Intelligent Systems and Technologies*.

[Aylett and Pidcock, 2007] Aylett, M. and Pidcock, C. (2007). The CereVoice characterful speech synthesiser SDK. In *Proceedings of AISB 2007*, pages 174–178, Newcastle, U.K.

[Benesty et al., 2007] Benesty, J., Sondhi, M. M., and Huang, Y. A. (2007). *Springer Handbook of Speech Processing*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Benoit et al., 1996] Benoit, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4):381–392.

[Bickley et al., 1997] Bickley, C., Stevens, K., and Williams, D. (1997). A framework for synthesis of segments based on pseudoarticulatory parameters. pages 211–220.

[Black and Campbell, 1995] Black, A. and Campbell, N. (1995). Optimising selection of units from speech database for concatenative synthesis. In *Proc. EUROSPEECH-95*, pages 581–584.

[Black et al., 1999] Black, A., Taylor, P., and Caley, R. (1999). *The Festival Speech Synthesis System*. University of Edinburgh.

[Black et al., 2007] Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proc. ICASSP 2007*, pages 1229–1232.

[Bodo, 2009] Bodo, A. Z. (2009). *Contribuţii la sinteza vorbirii în limba română*. PhD thesis, Technical University of Cluj-Napoca.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

[Burileanu et al., 1999] Burileanu, D., Sima, M., and Neagu, A. (1999). A phonetic converter for speech synthesis in romanian. In *Proceedings of the XIVth International Congress on Phonetic Sciences ICPhS99*.

[Buza, 2010] Buza, O. (2010). *Contribuţii la analiza şi sinteza vorbirii din text pentru limba română*. PhD thesis, Technical University of Cluj-Napoca.

[Calacean and Nivre, 2009] Calacean, M. and Nivre, J. (2009). A Data-Driven Dependency Parser for Romanian. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories.*, pages 65–76.

[Chitoran, 2002] Chitoran, I. (2002). *The phonology and morphology of Romanian diphtongization*. Probus.

[Cooke et al., 2006] Cooke, M., Barker, J., Cunningham, S., and X.Shao (2006). An audio-visual corpus for speech perception and automatic speech recognition. In *Journal of the Acoustical Society of America*, volume 20.

[Curteanu et al., 2007] Curteanu, N., Trandabat, D., and Moruz, M. (2007). Topic-Focus Articulation Algorithm on the Syntax-Prosody Interface of Romanian. In *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 516–523. Springer Berlin / Heidelberg.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

[D'Este and Bakker, 2010] D'Este, F. and Bakker, E. (2010). Articulatory Speech Synthesis with Parallel Multi-Objective Genetic Algorithms. In *Proceedings of ASCI*.

[DEX online-webpage, 2011] DEX online-webpage (2011). `http://dexonline.ro/download/dex-database.sql.gz`.

[Domokos et al., 2011] Domokos, J., Buza, O., and Toderean, G. (2011). Automated grapheme-to-phoneme conversion system for Romanian. In *Proceedings of the 6th Conference on Speech Technology and Human-Computer Dialogue*.

[Dudley, 1940] Dudley, H. (1940). The Carrier Nature of Speech. *The Bell System Technical Journal*.

[Dutoit et al., 1996] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and der Vrecken, O. V. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proc. ICSLP '96*, volume 3, pages 1393–1396, Philadelphia, PA.

[Eiben and Smith, 2010] Eiben, A. and Smith, J. (2010). *Introduction to Evolutionary Computing*. Springer.

[Falaschi et al., 1989] Falaschi, A., Giustiniani, M., and Verola, M. (1989). A hidden Markov model approach to speech synthesis. In *Proceedings of Eurospeech*, volume 1989, pages 2187–2190.

[Fant, 2005] Fant, G. (2005). *Speech Acoustics and Phonetics: Selected Writings*, chapter Speech Perception, pages 199–220. Springer Netherlands.

[Ferencz, 1997] Ferencz, A. (1997). *Contribuţii la dezvoltarea sintezei text-vorbire pentru limba română*. PhD thesis, Technical University of Cluj-Napoca.

[Frunză et al., 2005] Frunză O., Inkpen, D., and Nadeau, D. (2005). A text processing tool for the Romanian language. In *Proceedings of EuroLAN 2005: Workshop on Cross-Language Knowledge Induction*.

[Fujisaki and Ohno, 1998] Fujisaki, H. and Ohno, S. (1998). The use of a generative model of F0 contours for multilingual speech synthesis. In *Proceedings of ICSLP-1998*, pages 714–717.

[Fukumoto, 2010] Fukumoto, M. (2010). Interactive Evolutionary Computation Utilizing Subjective Evaluation and Physiological Information as Evaluation Value. In *Systems Man and Cybernetics*, pages 2874 – 2879.

[Geatbox-webpage, 2011] Geatbox-webpage (2011). `http://www.geatbx.com/docu/algindex-01.html`.

[Giurgiu, 2006] Giurgiu, M. (2006). Experimental Results on Prosody for Romanian Text to Speech Synthesis. *Revue Roumaine de Linguistique*, pages 467–476.

[Giurgiu and Peev, 2006] Giurgiu, M. and Peev, L. (2006). *Sinteza din text a semnalului vocal*. Risoprint.

[Gronnum, 1995] Gronnum, N. (1995). Superposition and subordination in intonation: a non-linear approach. In *Proceedings of the 13th International Congress of Phonetic Sciences*, volume 2, pages 124–131, Stockholm.

[Hansen, 2005] Hansen, N. (2005). The CMA evolution strategy: A tutorial. Technical report, TU Berlin, ETH Zurich.

[Hansen and Ostermeier, 1996] Hansen, N. and Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adap-

tation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, pages 312 –317.

[Hirst and Cristo, 1998] Hirst, D. and Cristo, A. D. (1998). *Intonation Systems: a survey of twenty languages*. Cambridge University Press.

[Holland, 1975] Holland, H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

[HTS webpage, 2010] HTS webpage (2010). `http://hts.sp.nitech.ac.jp/`.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall.

[Huda et al., 2009] Huda, S., Yearwood, J., and Togneri, R. (2009). A constraint-based evolutionary learning approach to the expectation maximization for optimal estimation of the hidden markov model for speech signal modeling. *Trans. Sys. Man Cyber. Part B*, 39:182–197.

[Hunt and Black, 1996] Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP-96*, pages 373–376.

[Imai et al., 1983] Imai, S., Sumita, K., and Furuichi, C. (1983). Mel log spectrum approximation (mlsa) filter for speech synthesis. *Electronics and Communications in Japan (Part I: Communications)*, 66(2):10–18.

[Jilka et al., 1999] Jilka, M., Mhler, G., and Dogil, G. (1999). Rules for the generation of tobi-based american english intonation. *Speech Communication*, 28:83–108.

[Jitcă et al., 2002] Jitcă D., Apopei, V., and Grigoraş F. (2002). Text-to-Speech System for Romanian Language based on Formantic Synthesis. In *European Conference on Intelligent Technologies*.

[Jitcă et al., 2008] Jitcă D., Apopei, V., and Jitcă M. (2008). A description language at the accentual unit level for Romanian intonation. In *In Proceedings "LangTech2008"*.

[Jong, 2006] Jong, K. D. (2006). *Evolutionary computation: a unified approach*. MIT Press, Cambridge MA.

[Kabir and Giurgiu, 2010] Kabir, A. and Giurgiu, M. (2010). A Romanian Corpus for Speech Perception and Automatic Speech Recognition. In *Recent Researches in Communications, Automation, Signal Processing, Nanotechnology, Astronomy and Nuclear Physics*.

[Karaiskos et al., 2008] Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard Challenge 2008. In *Proceedings of Blizzard Challenge Workshop*, Brisbane, Australia.

[Kawahara et al., 2001] Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *2nd MAVEBA*.

[Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207.

[Klatt, 1980] Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of The Acoustical Society of America*, 67.

[Lambert and Breen, 2004] Lambert, T. and Breen, A. P. (2004). A database design for a TTS synthesis system using lexical diphones. In *Proceedings of Interspeech*.

[Latorre and Akamine, 2008] Latorre, J. and Akamine, M. (2008). Multilevel Parametric-Base F0 Model for Speech Synthesis. In *Proceedings of Interspeech*.

[Leggetter and Woodland, 1995] Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. 9:171.

[Lv et al., 2009] Lv, S., Wang, S., and Wang, X. (2009). Emotional speech synthesis by XML file using interactive genetic algorithms. In *Proceedings of GEC Summit*, pages 907–910.

[Marques et al., 2010] Marques, V. M., Reis, C., and Machado, J. A. T. (2010). Interactive Evolutionary Computation in Music. In *Systems Man and Cybernetics*, pages 3501–3507.

[Matoušek et al., 2005] Matoušek, J., Hanzlíček, Z., and Tihelka, D. (2005). Hybrid syllable/triphone speech synthesis. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 2529–2532, Lisbon, Portugal.

[McDermott et al., 2010] McDermott, J., O'Neill, M., and Griffith, N. J. L. (2010). Interactive EC control of synthesized timbre. *Evolutionary Computation*, 18:277–303.

[Moisa et al., 2001] Moisa, T., Ontanu, D., and Dediu, A. (2001). Speech synthesis using neural networks trained by an evolutionary algorithm. In *Computational Science - ICCS 2001*, volume 2074 of *Lecture Notes in Computer Science*, pages 419–428. Springer Berlin / Heidelberg.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–468.

[Muralishankar et al., 2004] Muralishankar, R., Ramakrishnan, A., and Prathibha, P. (2004). Modification of pitch using DCT in the source domain. *Speech Communication*, 42(2):143 – 154.

[Muraoka et al., 1978] Muraoka, T., Yamada, Y., and Yamazaki, M. (1978). Sampling-frequency considerations in digital audio. *J. Audio Eng. Soc*, 26(4):252–256.

[Naseem et al., 2009] Naseem, T., Snyder, B., Eisenstein, J., and Barzilay, R. (2009). Multilingual part-of-speech tagging: two unsupervised approaches. *J. Artif. Int. Res.*, 36:341–385.

[Ohman, 1967] Ohman, S. (1967). *Word and sentence intonation.* STL-QPSR 2-3.

[Ohtani et al., 2006] Ohtani, Y., Toda, T., Saruwatari, H., and Shikano, K. (2006). Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proceedings of Interspeech 2006*, pages 2266–2269.

[Olive et al., 1993] Olive, J. P., Greenwood, A., and Coleman, J. (1993). *Acoustics of American English speech: a dynamic approach*. Springer.

[Ordean et al., 2009] Ordean, M., Saupe, A., Ordean, M., Duma, M., and Silaghi, G. (2009). Enhanced Rule-Based Phonetic Transcription for the Romanian Language. In *Proceedings of the 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, SYNASC '09, pages 401–406.

[Palmer, 1922] Palmer, H. (1922). *English Intonation with Systematic Exercises*. Cambridge University Press.

[Panait and Luke, 2002] Panait, L. and Luke, S. (2002). A comparison of two competitive fitness functions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, Proceedings of GECCO 2002, pages 503–511.

[Patterson, 1982] Patterson, R. (1982). Auditory filter shapes derived with noise stimuli. *Journal of the Acoustical Society of America*, 76:640–654.

[Qian et al., 2009] Qian, Y., Wu, Z., and Soong, F. (2009). Improved Prosody Generation by Maximizing Joint Likelihood of State and Longer Units. In *Proceedings of ICASSP*.

[Rutkowski, 2008] Rutkowski, L. (2008). *Computational Intelligence. Methods and Techniques*. Springer.

[Sabou et al., 2008] Sabou, O., Borza, P., and Tatar, D. (2008). POS Tagger for Romanian Language. `http://www.cs.ubbcluj.ro/~dtatar/nlp/WebTagger/WebTagger.htm`.

[Saito et al., 1996] Saito, T., Hashimoto, Y., and Sakamoto, M. (1996). High-quality speech synthesis using context-dependent syllabic units. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference - Volume 01*, ICASSP '96, pages 381–384.

[Sakai, 2005] Sakai, S. (2005). Additive modelling of English F0 contour for Speech Synthesis. In *Proceedings of ICASSP*.

[Santen et al.,] Santen, J. V., Mishra, T., and Klabbers, E. Estimating Phrase Curves in the General Superpositional Intonation Model. In *In Proceedings of the ISCA Speech Synthesis Workshop04*.

[Sato, 2005] Sato, Y. (2005). Voice quality conversion using interactive evolution of prosodic control. *Appl. Soft Comput.*, 5:181–192.

[Shinoda and Watanabe, 2000] Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Japan (E)*, 21:79–86.

[Shiohan et al., 2002] Shiohan, O., Myrvoll, T., and Lee, C. (2002). Structural maximum a posteriori linear regression for fast hmm adaptation. 16:5–24.

[Smith III and Abel, 1999] Smith III, J. O. and Abel, J. S. (1999). Bark and ERB bilinear transforms. *IEEE Trans. on Speech Audio Process.*, 7(6):697–708.

[Stan, 2010] Stan, A. (2010). Raport de cercetare ştiinţifică 1: Elaborarea şi dezvoltarea unui sistem de sinteză text-vorbire în limba română bazat pe modele Markov, independent de elementele de prozodie aferente textului.

[Stan, 2011a] Stan, A. (2011a). Raport de cercetare ştiinţifică 2: Elaborarea şi dezvoltarea unor metode deterministe de analiză şi control a prozodiei în limba română.

[Stan, 2011b] Stan, A. (2011b). Raport de cercetare ştiinţifică 3: Elaborarea şi dezvoltarea unor metode probabilistice de analiză şi control a prozodiei în limba română.

[Stan and Giurgiu, 2010] Stan, A. and Giurgiu, M. (2010). Romanian language statistics and resources for text-to-speech systems. In *Proceedings of ISETC 2010*, Timişoara, România.

[Stan and Giurgiu, 2011] Stan, A. and Giurgiu, M. (2011). A Superpositional Model Applied to F0 Parametrisation using DCT for Text-to-Speech Synthesis. In *Proceedings*

of the 6^{th} *Conference on Speech Technology and Human-Computer Dialogue*, Braşov, Romania.

[Stan et al., 2011a] Stan, A., Pop, F.-C., Cremene, M., Giurgiu, M., and Pallez, D. (2011a). Interactive Intonation Optimisation Using CMA-ES and DCT Parametrisation of the F0 Contour for Speech Synthesis. In *Proceedings of the* 5_{th} *Workshop on Nature Inspired Cooperative Strategies for Optimisation*, volume 387 of *Studies in Computational Intelligence*. Springer.

[Stan et al., 2011b] Stan, A., Yamagishi, J., King, S., and Aylett, M. (2011b). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, 53(3):442 – 450.

[Stan et al., 2011c] Stan, A., Yamagishi, J., King, S., and Aylett, M. (2011c). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, 53(3):442 – 450.

[Stylianou et al., 1998] Stylianou, Y., Cappé, O., and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Trans. on Speech Audio Process.*, 6(2):131–142.

[Sun, 2002] Sun, X. (2002). F0 generation for speech synthesis using a multi-tier approach. In *Proceedings of ICSLP*.

[Tao et al., 2006] Tao, J., Kang, Y., and Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Trans. on Audio Speech and Language Processing*, 14(4):1145–1154.

[Tatham and Morton, 2005] Tatham, M. and Morton, K. (2005). *Developments in speech synthesis*. John Wiley & Sons.

[Taylor, 2009] Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

[Teodorescu et al., 2010] Teodorescu, H.-N., Pistol, L., Feraru, M., Zbancioc, M., and Trandabat, D. (2010). Sounds of the Romanian Language Corpus. `http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm`.

[Teutenberg et al., 2008] Teutenberg, J., Wilson, C., and Riddle, P. (2008). Modelling and Synthesising F0 Contours with the Discrete Cosine Transform. In *Proceedings of ICASSP*.

[Toda and Tokuda, 2007] Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. & Syst.*, E90-D(5):816–824.

[Tokuda et al., 1991] Tokuda, K., Kobayashi, T., Fukada, T., Saito, H., and Imai, S. (1991). Spectral estimation of speech based on mel-cepstral representation. *IEICE Trans. Fundamentals*, J74-A(8):1240–1248. in Japanese.

[Tokuda et al., 1994a] Tokuda, K., Kobayashi, T., and Imai, S. (1994a). Recursive calculation of mel-cepstrum from LP coefficients. In *Technical Report of Nagoya Institute of Technology*.

[Tokuda et al., 1994b] Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994b). Mel-generalized cepstral analysis – a unified approach to speech spectral estimation. In *Proc. ICSLP-94*, pages 1043–1046, Yokohama, Japan.

[Tokuda et al., 1999] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proc. ICASSP-99*, pages 229–232.

[Tokuda et al., 2002a] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002a). Multi-space probability distribution HMM. *IEICE Trans. Inf. & Syst.*, E85-D(3):455–464.

[Tokuda et al., 2002b] Tokuda, K., Zen, H., and Black, A. (2002b). An HMM-based speech synthesis system applied to English. In *Proc. IEEE Speech Synthesis Workshop*.

[Toma and Munteanu, 2009] Toma, S.-A. and Munteanu, D.-P. (2009). Rule-Based Automatic Phonetic Transcription for the Romanian Language. In *Proceedings of the 2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns*, pages 682–686.

[Tufis et al., 2008] Tufis, D., Irimia, E., Ion, R., and Ceausu, A. (2008). Unsupervised lexical acquisition for part of speech tagging. In *LREC*. European Language Resources Association.

[Vaseghi, 2007] Vaseghi, S. V. (2007). *Multimedia Signal Processing*. Wiley & Sons.

[Vlad and Mitrea, 2002] Vlad, A. and Mitrea, A. (2002). *Contribuţii privind structura statistică de cuvinte în limba română scrisă*. Ed Expert, Bucharest.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Sys. Morgan Kaufmann, second edition.

[Wolf, 1981] Wolf, H. (1981). Control of prosodic parameters for a formant synthesizer based on diphone concatenation. In *IEEE International Conference Acoustics, Speech, and Signal Processing*, page 106.

[Wu et al., 2008] Wu, Z., Qian, Y., Soong, F., and Zhang, B. (2008). Modeling and Generating Tone Contour with phrase Intonation for Mandarin Chinese Speech. In *Proceedings of ISCSLP*.

[Yamagishi, 2006] Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis*. PhD thesis, Tokyo Institute of Technology, Tokyo.

[Yamagishi and King, 2010] Yamagishi, J. and King, S. (2010). Simple methods for improving speaker-similarity of HMM-based speech synthesis. In *Proc. ICASSP 2010*, pages 4610–4613, Dallas, TX.

[Yamagishi et al., 2009] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Audio, Speech, & Language Processing*, 17(1):66–83.

[Yamagishi et al., 2008a] Yamagishi, J., Ling, Z., and King, S. (2008a). Robustness of HMM-based speech synthesis. In *Proc. Interspeech 2008*, pages 581–584, Brisbane, Australia.

[Yamagishi et al., 2005] Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. *IEICE - Trans. Inf. Syst.*, E88-D:502–509.

[Yamagishi et al., 2008b] Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008b). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008*, Brisbane, Australia.

[Young et al., 2001] Young, S., Everman, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2001). *The HTK Book Version 3.1.*

[Young et al., 1994] Young, S., Odell, J., and Woodland, P. (1994). Tree-based state tying for high accuracy acoustic modeling. In *Proc. ARPA Human Language Technology Workshop*, pages 307–312.

[Zen et al., 2007a] Zen, H., Nose, T., Yamagishi, J., Sako, S., and Tokuda, K. (2007a). The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of the Sixth ISCA Workshop on Speech Synthesis*, pages 294–299.

[Zen et al., 2007b] Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007b). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. & Syst.*, E90-D(1):325–333.

[Zen et al., 2009] Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

[Zen et al., 2007c] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007c). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. & Syst.*, E90-D(5):825–834.

[Zwicker and Scharf, 1965] Zwicker, E. and Scharf, B. (1965). A model of loudness summation. *Psych. Rev.*, 72:2–26.

Appendix A

List of the Phonemes Used in the Speech Synthesiser

Phoneme	Sample word	SAMPA notation
a	m-*a*-ria	a
@	cas-*ă*-	@
a@	m-*â*-nă /-*î*-nceput	1
b	a-*b*-ac	b
k	a-*c*-t	k
ch	a-*ch*-eea	tS
d	-*d*-acă	d
e	d-*e*-spre	e
e@	c-*e*-as	e_X
f	-*f*-apt	f
g	a-*g*-onie	g
dz	-*g*-eam	dZ
h	-*h*-artă	h
i	in-*i*-mă	i
j	-*i*-epure	j
ij	câin-*i*	i_0
zh	a-*j*-utor	Z
l	a-*l*-ta	l
m	a-*m*-ară	m
n	î-*n*-să	n
o	m-*o*-tor	o
o@	-*o*-aie	o_X
p	a-*p*-ă	p
r	a-*r*-tă	r
s	a-*s*-ta	s
sh	-*ş*-i	S
t	ta-*t*-a	t
ts	-*ţ*-ară	ts
u	m-*u*-nca	u
w	plo-*u*-at	w
v	-*v*-ara	v
z	a-*z*-i	z

Appendix B

Letter-to-Sound Rules Written in Festival

```
(lts.ruleset
    romanian
    ;;
    ;;;;;;;;;Sets used;;;;;;;;;;;;;;;;
    (
    ;; For the pronunciation of "ce", "ci", "ge", "gi", "che", "chi", "ghe", "ghi".
    ( CV a ă â o u b c d f g j k l m n p r s şt ţu v z )
    ;; Consonants
    ( C1 b c d f g j k l m n p r s ş t ţ u v z )
    ;;Vowels
    ( V1 a ă â e i î o u )
    ;;B followed by t k ţ f s ş h
    ( SUR t k ţ f s ş h )
    );;end sets
    ;;
    ;;;;;;;;;;;;Rules;;;;;;;;;;;;;;;;;
    (
    ( [ ă ] = @ )
    ;;"X" followed by consonant, preceeded by "e"
    ( e [ x ] C1 = k s );;extemporal
    ;;"X" followed by vowel, preceeded by "e"
    ( e [ x ] V1 = g z );; examen
    ;;"X" preceeded by "a", "o"
    ( a [ x ] = k s );;axat, oxida
    ;; Pronunciation of î, Î is the same as for â,
    ( [ î ] = a@ )
    ;;Groups: "ce", "ci", "ge", "gi", "che", "chi", "ghe", "ghi"
    ( [ c ] CV = k )
    ( [ c ] e = ch )
    ( [ c ] i = ch )
    ( [ c h ] i CV = k )
    ( [ c h ] e CV = k )
    ( [ g ] CV = g )
    ( [ g ] e = dz )
```

```
    ( [ g ] i = dz )
    ( [ g h ] i CV = g )
    ( [ g h ] e CV = g )
    ;;"b" followed by SUR = p:
    ( [ b ] SUR = p );;"subtil"<=>"suptil"
    ;;Total asimilation: subpământean<=>supământean
    ( [ b p ] = p )
    ;;"E" in initial position - some cases: pronouns eu, ea, el, ei, ele and some forms of the verb
"to be" = "a fi"
    ( # [ e r a u ] # = j e r a w );;erau->ierau
    ( # [ e s t e ] # = j e s t e );;este->ieste
    ( # [ e r a ţ i ] # = j e r a ts ij );;
    ( # [ e r a m ] # = j e r a m )
    ( # [ e ] # = j e )
    ( # [ e ş t i ] # = j e sh t ij )
    ( # [ e r a ] # = j e r a )
    ( # [ e r a i ] # = j e r a j )
    ( # [ e u ] # = j e w )
    ( # [ e a ] # = j a )
    ( # [ e i ] # = j e j )
    ( # [ e l ] # = j e l )
    ( # [ e l e ] # = j e l e )
    ;;Diphtongs - falling
    ( [ a i ] = a j );;rai
    ( [ a u ] = a w );;sau
    ( [ e i ] = e j );;trei
    ( [ e u ] = e w );;greu
    ( [ i i ] = i j );;mii
    ( [ i u ] = i w );;scriu
    ( [ o i ] = o j );;noi
    ( [ o u ] = o w );;bou
    ( [ u i ] = u j );;pui
    ( [ ă i ] = @ j );;răi
    ( [ ă u ] = @ w );;dulău
    ( [ â i ] = a@ j );;câine
    ( [ â u ] = a@ w );;râu
    ( [ u u ] = u w );;continuu
    ;;Diphtongs - rising
    ( [ e a ] = e@ a );;beată
    ( [ e o ] = e@ o );;Gheorghe
    ( [ i a ] = j a );;biata
    ( [ i e ] = j e );;fier
    ( [ i o ] = j o );;iod
    ( [ i u ] = j u );;iubit
    ( [ o a ] = o@ a );;găoace
    ( [ u e ] = w e );;piuez
    ( [ u a ] = w a );;băcăuan
    ( [ u ă ] = w @ );;două
    ( [ u â ] = w a@ );;plouând
    ;;Triphtongs
```

```
( [ e a i ] = e@ a j ) ;;ceainic
( [ e a u ] = e@ a w );;beau
( [ i a i ] = j a j ) ;;mi-ai
( [ i a u ] = j a w ) ;;suiau
( [ i e i ] = j e j ) ;;piei
( [ i e u ] = j e w ) ;;maieu
( [ i o i ] = j o j ) ;;picioică
( [ i o u ] = j o w ) ;;maiou
( [ o a i ] = o@ a j ) ;;leoaică
( [ u a i ] = w a j ) ;;nşeuai
( [ u a u ] = w a w ) ;;nşeuau
( [ u ă i ] = w @ j ) ;;rouăı
( [ e o a ] = e@ o@ a ) ;;pleoape
( [ i o a ] = j o@ a ) ;;creioane
;; Final position for "i" not a diphtong
( C1 [ i ] # = ij ) ;;câini
;;Vowels
( [ a ] = a )
( [ e ] = e )
( [ i ] = i )
( [ i ] = ij )
( [ i ] = j )
( [ o ] = o )
( [ o ] = o@ )
( [ u ] = u )
( [ u ] = w )
( [ ă ] = @ )
( [ Ă ] = @ )
( [ â ] = a@ )
( [ Â ] = a@ )
( [ î ] = a@ )
( [ Î ] = a@ )
;;Consonants
( [ c ] = k )
( [ b ] = b )
( [ d ] = d )
( [ f ] = f )
( [ g ] = g )
( [ h ] = h )
( [ j ] = zh )
( [ k ] = k )
( [ l ] = l )
( [ m ] = m )
( [ n ] = n )
( [ p ] = p )
( [ q ] = k )
( [ r ] = r )
( [ s ] = s )
( [ ş ] = sh )
( [ Ş ] = sh )
```

```
( [ t ] = t )
( [ ţ ] = ts )
( [ Ţ ] = ts )
( [ v ] = v )
( [ w ] = u )
( [ x ] = k s )
( [ y ] = i )
( [ z ] = z )
( # [ c ] # = ch e )
( # [ d ] # = d e )
( # [ f ] # = e f )
( # [ b ] # = b e )
( # [ g ] # = dz e )
( # [ h ] # = h a sh ij )
( # [ j ] # = zh a@ )
( # [ k ] # = k a )
( # [ l ] # = e l )
( # [ m ] # = e m )
( # [ n ] # = e n )
( # [ p ] # = p e )
( # [ q ] # = k i w )
( # [ r ] # = e r )
( # [ s ] # = e s )
( # [ t ] # = t e )
( # [ v ] # = v e )
( # [ x ] # = i c s )
( # [ y ] # = i g r e k )
( # [ z ] # = z e t )
( # [ ş ] # = sh a@ )
( # [ ţ ] # = ts a@ )
))
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
```

Appendix C

Sample Entries of the 65,000 Word Lexicon

An extract from the lexicon. The list of phonemes is presented in Appendix A, and 0 and 1 represent the accent [1]. Only one accent can exist within a word.

a	a1
ă	@1
â	a@1
aa	a1_ a0
ab	a1_ b
aba	a0_ b_ a1
abajururile	a0_ b_ a0_ zh_ u1_ r_ u0_ r_ i0_ l_ e0
abandon	a0_ b_ a0_ n_ d_ o1_ n
abandona	a0_ b_ a0_ n_ d_ o0_ n_ a1
abandonând	a0_ b_ a0_ n_ d_ o0_ n_ a@1_ n_ d
abandonarea	a0_ b_ a0_ n_ d_ o0_ n_ a1_ r_ e@0_ a0
abandonării	a0_ b_ a0_ n_ d_ o0_ n_ @1_ r_ i0_ j0
abandonat	a0_ b_ a0_ n_ d_ o0_ n_ a1_ t
abandonată	a0_ b_ a0_ n_ d_ o0_ n_ a1_ t_ @0
abandonate	a0_ b_ a0_ n_ d_ o0_ n_ a1_ t_ e0
abandonaţi	a0_ b_ a0_ n_ d_ o0_ n_ a1_ ts_ ij0
abandonează	a0_ b_ a0_ n_ d_ o0_ n_ e@0_ a1_ z_ @0
abandoneze	a0_ b_ a0_ n_ d_ o0_ n_ e1_ z_ e0
abandonul	a0_ b_ a0_ n_ d_ o1_ n_ u0_ l
abandonului	a0_ b_ a0_ n_ d_ o1_ n_ u0_ l_ u0_ j0
abat	a0_ b_ a1_ t
abată	a0_ b_ a1_ t_ @0
abate	a0_ b_ a1_ t_ e0
abatere	a0_ b_ a1_ t_ e0_ r_ e0
abaterea	a0_ b_ a1_ t_ e0_ r_ e@0_ a0
abateri	a0_ b_ a1_ t_ e0_ r_ ij0
abaterile	a0_ b_ a1_ t_ e0_ r_ i0_ l_ e0
abator	a0_ b_ a0_ t_ o1_ r

[1] 0-no accent, 1-accent

abatorul	a0_ b_ a0_ t_ o1_ r_ u0_ l
abatorului	a0_ b_ a0_ t_ o1_ r_ u0_ l_ u0_ j0
abătut	a0_ b_ @0_ t_ u1_ t
abdică	a0_ b_ d_ i0_ k_ @1
abdice	a0_ b_ d_ i1_ ch_ e0
abdomen	a0_ b_ d_ o0_ m_ e1_ n
abdomenul	a0_ b_ d_ o0_ m_ e1_ n_ u0_ l
abdomenului	a0_ b_ d_ o0_ m_ e1_ n_ u0_ l_ u0_ j0
abdominală	a0_ b_ d_ o0_ m_ i0_ n_ a1_ l_ @0
abdominale	a0_ b_ d_ o0_ m_ i0_ n_ a1_ l_ e0
abecedar	a0_ b_ e0_ ch_ e0_ d_ a1_ r
abecedarul	a0_ b_ e0_ ch_ e0_ d_ a1_ r_ u0_ l
abecedarului	a0_ b_ e0_ ch_ e0_ d_ a1_ r_ u0_ l_ u0_ j0
aberant	a0_ b_ e0_ r_ a1_ n_ t
aberantă	a0_ b_ e0_ r_ a1_ n_ t_ @0
aberante	a0_ b_ e0_ r_ a1_ n_ t_ e0
aberaţie	a0_ b_ e0_ r_ a1_ ts_ i0_ e0
aberaţii	a0_ b_ e0_ r_ a1_ ts_ i0_ j0
abia	a0_ b_ j0_ a1
abil	a0_ b_ i1_ l
abilitare	a0_ b_ i0_ l_ i0_ t_ a1_ r_ e0
abilitarea	a0_ b_ i0_ l_ i0_ t_ a1_ r_ e@0_ a0
abilitat	a0_ b_ i0_ l_ i0_ t_ a1_ t
abilitate	a0_ b_ i0_ l_ i0_ t_ a1_ t_ e0
abilitatea	a0_ b_ i0_ l_ i0_ t_ a1_ t_ e@0_ a0
abilitaţi	a0_ b_ i0_ l_ i0_ t_ a1_ ts_ ij0
abilităţi	a0_ b_ i0_ l_ i0_ t_ @1_ ts_ ij0
abilităţile	a0_ b_ i0_ l_ i0_ t_ @1_ ts_ i0_ l_ e0
abilităţilor	a0_ b_ i0_ l_ i0_ t_ @1_ ts_ i0_ l_ o0_ r
abilitau	a0_ b_ i0_ l_ i0_ t_ a1_ w0
abilitează	a0_ b_ i0_ l_ i0_ t_ e@0_ a1_ z_ @0
abisului	a0_ b_ i1_ s_ u0_ l_ u0_ j0
abitir	a0_ b_ i0_ t_ i1_ r
abject	a0_ b_ zh_ e1_ k_ t
abolirea	a0_ b_ o0_ l_ i1_ r_ e@0_ a0
abolit	a0_ b_ o0_ l_ i1_ t
abolite	a0_ b_ o0_ l_ i1_ t_ e0
abona	a0_ b_ o0_ n_ a1
abonament	a0_ b_ o0_ n_ a0_ m_ e1_ n_ t
abonamente	a0_ b_ o0_ n_ a0_ m_ e1_ n_ t_ e0
abonamentele	a0_ b_ o0_ n_ a0_ m_ e1_ n_ t_ e0_ l_ e0
abonamentelor	a0_ b_ o0_ n_ a0_ m_ e1_ n_ t_ e0_ l_ o0_ r
abonamentul	a0_ b_ o0_ n_ a0_ m_ e1_ n_ t_ u0_ l
abonamentului	a0_ b_ o0_ n_ a0_ m_ e1_ n_ t_ u0_ l_ u0_ j0
abonat	a0_ b_ o0_ n_ a1_ t

abonată	a0_ b_ o0_ n_ a1_ t_ @0
abonaţi	a0_ b_ o0_ n_ a1_ ts_ ij0
abonatul	a0_ b_ o0_ n_ a1_ t_ u0_ l
aborda	a0_ b_ o0_ r_ d_ a1
abordăm	a0_ b_ o0_ r_ d_ @1_ m
abordare	a0_ b_ o0_ r_ d_ a1_ r_ e0
abordare	a0_ b_ o0_ r_ d_ a1_ r_ e0
abordarea	a0_ b_ o0_ r_ d_ a1_ r_ e@0_ a0
abordări	a0_ b_ o0_ r_ d_ @1_ r_ ij0
abordării	a0_ b_ o0_ r_ d_ @1_ r_ i0_ j0
abordările	a0_ b_ o0_ r_ d_ @1_ r_ i0_ l_ e0
abordasem	a0_ b_ o0_ r_ d_ a1_ s_ e0_ m
abordat	a0_ b_ o0_ r_ d_ a1_ t
abordată	a0_ b_ o0_ r_ d_ a1_ t_ @0
abordate	a0_ b_ o0_ r_ d_ a1_ t_ e0
abordaţi	a0_ b_ o0_ r_ d_ a1_ ts_ ij0
abordează	a0_ b_ o0_ r_ d_ e@0_ a1_ z_ @0
abordeze	a0_ b_ o0_ r_ d_ e1_ z_ e0
abordezi	a0_ b_ o0_ r_ d_ e1_ z_ ij0

Appendix D

HTS Labels Format

p_1~p_2-p_3+p_4=p_5:p_6_p_7
/A/a_1_a_2_a_3
/B/b_1 - b_2 - b_3 : b_4 - b_5 & b_6 - b_7 # b_8 - b_9 $ b_{10} - b_{11} > b_{12} - b_{13} < b_{14} - b_{15} — b_{16}
/C/c_1+c_2+c_3
/D/d_1_ d_2
/E/e_1+e_2:e_3+e_4&e_5+e_6#e_7+e_8
/F/f_1_ f_2
/G/g_1_ g_2
/H/h_1=h_2:h_3=h_4&h_5
/I/i_1_ i_2
/J/j_1+j_2-j_3

p_1	the phoneme identity before the previous phoneme
p_2	the previous phoneme identity
p_3	the current phoneme identity
p_4	the next phoneme identity
p_5	the phoneme identity after the next phoneme
p_6	position of the current phoneme identity in the current syllable (forward)
p_7	position of the current phoneme identity in the current syllable (backward)
a_1	whether the previous syllable is stressed or not (0:not stressed, 1:stressed)
a_2	whether the previous syllable is accented or not (0:not accented, 1:accented)
a_3	the number of phonemes in the previous syllable
b_1	whether the current syllable is stressed or not (0:not stressed, 1:stressed)
b_2	whether the current syllable is accented or not (0:not accented, 1:accented)
b_3	the number of phonemes in current syllable
b_4	position of the current syllable in the current word (forward)
b_5	position of the current syllable in the current word (backward)
b_6	position of the current syllable in the current phrase (forward)
b_7	position of the current syllable in the current phrase (backward)
b_8	the number of stressed syllables before the current syllable in the current phrase
b_9	the number of stressed syllables after the current syllable in the current phrase
b_{10}	the number of accented syllables before the current syllable in the current phrase
b_{11}	the number of accented syllables after the current syllable in the current phrase

b_{12}	the number of syllables from the previous stressed syllable to the current syllable
b_{13}	the number of syllables from the current syllable to the next stressed syllable
b_{14}	the number of syllables from the previous accented syllable to the current syllable
b_{15}	the number of syllables from the current syllable to the next accented syllable
b_{16}	name of the vowel of the current syllable
c_1	whether the next syllable is stressed or not (0:not stressed, 1:stressed
c_2	whether the next syllable is accented or not (0:not accented, 1:accented)
c_3	the number of phonemes in the next syllable
d_1	part-of-speech of the previous word
d_2	the number of syllables in the previous word
e_1	part-of-speech of the current word
e_2	the number of syllables in the current word
e_3	position of the current word in the current phrase (forward)
e_4	position of the current word in the current phrase (backward)
e_5	the number of content words before the current word in the current phrase
e_6	the number of content words after the current word in the current phrase
e_7	the number of words from the previous content word to the current word
e_8	the number of words from the current word to the next content word
f_1	part-of-speech of the next word
f_2	the number of syllables in the next word
g_1	the number of syllables in the previous phrase
g_2	the number of words in the previous phrase
h_1	the number of syllables in the current phrase
h_2	the number of words in the current phrase
h_3	position of the current phrase in the utterance (forward)
h_4	position of the current phrase in the utterance (backward)
h_5	TOBI endtone of the current phrase
i_1	the number of syllables in the next phrase
i_2	the number of words in the next phrase
j_1	the number of syllables in this utterance
j_2	the number of words in this utterance
j_3	the number of phrases in this utterance

Additionaly in the training labels, the temporal markers are added at the beginning of each line.

Sample line from a training label:

```
2700000 3650000 d~e-a+s=e:1_1
/A/1_1_2
/B/0-0-1:1-4& 2-4#1-1$ 1-1 > 0-0 < 0-0—a
/C/1+1+2
/D/content_1
/E/feature+4:2+1&1+0#0+0
/F/content_3
/G/0_0
/H/5=2:1=2&L-L%
/I/18_9
/J/22+11-2
```

Appendix E

HTS Label File Example

xx~xx-#+u=n:xx_xx/A/0_0_0/B/xx-xx-xx:xx-xx&xx-xx#xx-xx$xx-xx>xx-xx<xx-xx—xx/C/0+0+2/D/feature_0/E/xx+xx:xx+xx&xx+xx#xx+xx/F/content_3/G/0_0/H/xx=xx:1=4&L-L%/I/1_1/J/23+15-4
xx~#-u+n=m:2_1/A/0_0_0/B/1-1-2:1-1&1-12#0-5$0-5>0-2<0-2—u/C/0+0+2/D/feature_0/E/feature+1:1+6&0+4#0+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
~u-n+m=i:3_0/A/0_0_0/B/1-1-2:1-1&1-12#0-5$0-5>0-2<0-2—u/C/0+0+2/D/feature_0/E/feature+1:1+6&0+4#0+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
u~n-m+i=l:1_2/A/1_1_2/B/0-0-2:1-3&2-11#1-5$1-5>0-1<0-1—i/C/0+0+2/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
n~m-i+l=i:2_1/A/1_1_2/B/0-0-2:1-3&2-11#1-5$1-5>0-1<0-1—i/C/0+0+2/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
m~i-l+i=t:1_2/A/0_0_2/B/0-0-2:2-2&3-10#1-5$1-5>1-0<1-0—i/C/1+1+3/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
i~l-i+t=a:2_1/A/0_0_2/B/0-0-2:2-2&3-10#1-5$1-5>1-0<1-0—i/C/1+1+3/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
l~i-t+a=r:1_3/A/0_0_2/B/1-1-3:3-1&4-9#1-4$1-4>2-1<2-1—a/C/0+0+2/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
i~t-a+r=r:2_2/A/0_0_2/B/1-1-3:3-1&4-9#1-4$1-4>2-1<2-1—a/C/0+0+2/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
t~a-r+r=o:3_1/A/0_0_2/B/1-1-3:3-1&4-9#1-4$1-4>2-1<2-1—a/C/0+0+2/D/feature_1/E/content+3:2+5&0+3#1+0/F/content_3/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
a~r-r+o=m:1_2/A/1_1_3/B/0-0-2:1-3&5-8#2-4$2-4>0-0<0-0—o/C/1+1+2/D/content_3/E/content+3:3+4&1+2#0+1/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
r~r-o+m=a@:2_1/A/1_1_3/B/0-0-2:1-3&5-8#2-4$2-4>0-0<0-0—o/C/1+1+2/D/content_3/E/content+3:3+4&1+2#0+1/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
r~o-m+a@=n:1_2/A/0_0_2/B/1-1-2:2-2&6-7#2-3$2-3>1-1<1-1—a@/C/0+0+1/D/content_3/E/content+3:3+4&1+2#0+1/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
o~m-a@+n=a:2_1/A/0_0_2/B/1-1-2:2-2&6-7#2-3$2-3>1-1<1-1—a@/C/0+0+1/D/content_3/E/content+3:3+4&1+2#0+1/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
m~a@-n+a=f:1_1/A/1_1_2/B/0-0-1:3-1&7-6#3-3$3-3>0-0<0-0—no_vowels/C/1+1+1/D/content_3/E/content+3:3+4&1+2#0+1/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4

a@~n-a+f=o:1_1/A/0_0_1/B/1-1-1:1-1&8-5#3-2$3-2>1-0<1-0—a/C/1+1+3/D/content_3
/E/feature+1:4+3&2+2#0+0/F/content_2/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
n~a-f+o=s:1_3/A/1_1_1/B/1-1-3:1-2&9-4#4-1$4-1>0-2<0-2—o/C/0+0+1/D/feature_1
/E/content+2:5+2&2+1#1+0/F/content_2/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
a~f-o+s=t:2_2/A/1_1_1/B/1-1-3:1-2&9-4#4-1$4-1>0-2<0-2—o/C/0+0+1/D/feature_1
/E/content+2:5+2&2+1#1+0/F/content_2/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
f~o-s+t=u:3_1/A/1_1_1/B/1-1-3:1-2&9-4#4-1$4-1>0-2<0-2—o/C/0+0+1/D/feature_1
/E/content+2:5+2&2+1#1+0/F/content_2/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
o~s-t+u=ch:1_1/A/1_1_3/B/0-0-1:2-1&10-3#5-1$5-1>0-1<0-1—no_vowels/C/0+0+1
/D/feature_1/E/content+2:5+2&2+1#1+0/F/content_2/G/0_0/H/12=6:1=4&L-L%/I/1_1
/J/23+15-4
s~t-u+ch=i:1_1/A/0_0_1/B/0-0-1:1-2&11-2#5-1$5-1>1-0<1-0—u/C/1+1+3/D/content_2
/E/content+2:6+1&3+0#0+0/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
t~u-ch+i=s:1_3/A/0_0_1/B/1-1-3:2-1&12-1#5-0$5-0>2-0<2-0—i/C/1+1+5/D/content_2
/E/content+2:6+1&3+0#0+0/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
u~ch-i+s=pau:2_2/A/0_0_1/B/1-1-3:2-1&12-1#5-0$5-0>2-0<2-0—i/C/1+1+5/D/content_2
/E/content+2:6+1&3+0#0+0/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
ch~i-s+pau=m:3_1/A/0_0_1/B/1-1-3:2-1&12-1#5-0$5-0>2-0<2-0—i/C/1+1+5/D/content_2
/E/content+2:6+1&3+0#0+0/F/feature_1/G/0_0/H/12=6:1=4&L-L%/I/1_1/J/23+15-4
i~s-pau+m=a:xx_xx/A/0_0_1/B/xx-xx-xx:xx-xx&xx-xx#xx-xx$xx-xx>xx-xx<xx-xx—xx
/C/1+1+5/D/content_2/E/xx+xx:xx+xx&xx+xx#xx+xx/F/feature_1/G/0_0/H/xx=xx
:1=4&L-L%/I/1_1/J/23+15-4
s~pau-m+a=r:1_5/A/1_1_3/B/1-1-5:1-1&1-1#0-0$0-0>0-0<0-0—ij/C/1+1+1/D/content_2
/E/feature+1:1+1&0+0#0+0/F/content_2/G/12_6/H/1=1:2=3&L-L%/I/6_2/J/23+15-4
pau~m-a+r=ts:2_4/A/1_1_3/B/1-1-5:1-1&1-1#0-0$0-0>0-0<0-0—ij/C/1+1+1/D/content_2
/E/feature+1:1+1&0+0#0+0/F/content_2/G/12_6/H/1=1:2=3&L-L%/I/6_2/J/23+15-4
m~a-r+ts=ij:3_3/A/1_1_3/B/1-1-5:1-1&1-1#0-0$0-0>0-0<0-0—ij/C/1+1+1/D/content_2
/E/feature+1:1+1&0+0#0+0/F/content_2/G/12_6/H/1=1:2=3&L-L%/I/6_2/J/23+15-4
a~r-ts+ij=pau:4_2/A/1_1_3/B/1-1-5:1-1&1-1#0-0$0-0>0-0<0-0—ij/C/1+1+1/D/content_2
/E/feature+1:1+1&0+0#0+0/F/content_2/G/12_6/H/1=1:2=3&L-L%/I/6_2/J/23+15-4
r~ts-ij+pau=a@:5_1/A/1_1_3/B/1-1-5:1-1&1-1#0-0$0-0>0-0<0-0—ij/C/1+1+1/D/content_2
/E/feature+1:1+1&0+0#0+0/F/content_2/G/12_6/H/1=1:2=3&L-L%/I/6_2/J/23+15-4
ts~ij-pau+a@=n:xx_xx/A/1_1_3/B/xx-xx-xx:xx-xx&xx-xx#xx-xx$xx-xx>xx-xx<xx-xx—xx
/C/1+1+1/D/content_2/E/xx+xx:xx+xx&xx+xx#xx+xx/F/content_2/G/12_6/H/xx=xx:
2=3&L-L%/I/6_2/J/23+15-4
ij~pau-a@+n=a:1_1/A/1_1_5/B/1-1-1:1-2&1-6#0-1$0-1>0-3<0-3—a@/C/0+0+1/D/feature_1
/E/content+2:1+2&0+1#0+0/F/content_4/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
pau~a@-n+a=f:1_1/A/1_1_1/B/0-0-1:2-1&2-5#1-1$1-1>0-2<0-2—no_vowels/C/0+0+2
/D/feature_1/E/content+2:1+2&0+1#0+0/F/content_4/G/1_1/H/6=2:3=2&L-L%/I/11_6
/J/23+15-4
a@~n-a+f=g:1_2/A/0_0_1/B/0-0-2:1-4&3-4#1-1$1-1>1-1<1-1—a/C/0+0+2/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
n~a-f+g=a:2_1/A/0_0_1/B/0-0-2:1-4&3-4#1-1$1-1>1-1<1-1—a/C/0+0+2/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
a~f-g+a=n:1_2/A/0_0_2/B/0-0-2:2-3&4-3#1-1$1-1>2-0<2-0—a/C/1+1+2/D/content_2

/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
f~g-a+n=i:2_1/A/0_0_2/B/0-0-2:2-3&4-3#1-1$1-1>2-0<2-0—a/C/1+1+2
/D/content_2/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6
/J/23+15-4
g~a-n+i=s:1_2/A/0_0_2/B/1-1-2:3-2&5-2#1-0$1-0>3-1<3-1—i/C/0+0+4/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
a~n-i+s=t:2_1/A/0_0_2/B/1-1-2:3-2&5-2#1-0$1-0>3-1<3-1—i/C/0+0+4/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
n~i-s+t=a:1_4/A/1_1_2/B/0-0-4:4-1&6-1#2-0$2-0>0-0<0-0—a/C/1+1+2/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
i~s-t+a=n:2_3/A/1_1_2/B/0-0-4:4-1&6-1#2-0$2-0>0-0<0-0—a/C/1+1+2/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
s~t-a+n=pau:3_2/A/1_1_2/B/0-0-4:4-1&6-1#2-0$2-0>0-0<0-0—a/C/1+1+2/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
t~a-n+pau=j:4_1/A/1_1_2/B/0-0-4:4-1&6-1#2-0$2-0>0-0<0-0—a/C/1+1+2/D/content_2
/E/content+4:2+1&1+0#0+0/F/feature_2/G/1_1/H/6=2:3=2&L-L%/I/11_6/J/23+15-4
a~n-pau+j=a:xx_xx/A/1_1_2/B/xx-xx-xx:xx-xx&xx-xx#xx-xx$xx-xx>xx-xx<xx-xx—xx
/C/1+1+2/D/content_2/E/xx+xx:xx+xx&xx+xx#xx+xx/F/feature_2/G/1_1/H/xx=xx:
3=2&L-L%/I/11_6/J/23+15-4
n~pau-j+a=r:1_2/A/0_0_4/B/1-1-2:1-2&1-11#0-5$0-5>0-1<0-1—a/C/0+0+1/D/content_4
/E/feature+2:1+6&0+4#0+1/F/feature_2/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
pau~j-a+r=u:2_1/A/0_0_4/B/1-1-2:1-2&1-11#0-5$0-5>0-1<0-1—a/C/0+0+1/D/content_4
/E/feature+2:1+6&0+4#0+1/F/feature_2/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
j~a-r+u=n:1_1/A/1_1_2/B/0-0-1:2-1&2-10#1-5$1-5>0-0<0-0—no_vowels/C/1+1+1
/D/content_4/E/feature+2:1+6&0+4#0+1/F/feature_2/G/6_2/H/11=6:4=1&L-L%/I/0_0
/J/23+15-4
a~r-u+n=a:1_1/A/0_0_1/B/1-1-1:1-2&3-9#1-4$1-4>1-1<1-1—u/C/0+0+1/D/feature_2
/E/feature+2:2+5&0+4#1+0/F/content_3/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
r~u-n+a=l:1_1/A/1_1_1/B/0-0-1:2-1&4-8#2-4$2-4>0-0<0-0—no_vowels/C/1+1+2
/D/feature_2/E/feature+2:2+5&0+4#1+0/F/content_3/G/6_2/H/11=6:4=1&L-L%/I/0_0
/J/23+15-4
u~n-a+l=t:1_2/A/0_0_1/B/1-1-2:1-3&5-7#2-3$2-3>1-2<1-2—a/C/0+0+2/D/feature_2
/E/content+3:3+4&0+3#2+0/F/content_1/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
n~a-l+t=u:2_1/A/0_0_1/B/1-1-2:1-3&5-7#2-3$2-3>1-2<1-2—a/C/0+0+2/D/feature_2
/E/content+3:3+4&0+3#2+0/F/content_1/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
a~l-t+u=l:1_2/A/1_1_2/B/0-0-2:2-2&6-6#3-3$3-3>0-1<0-1—u/C/0+0+1/D/feature_2
/E/content+3:3+4&0+3#2+0/F/content_1/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
l~t-u+l=a:2_1/A/1_1_2/B/0-0-2:2-2&6-6#3-3$3-3>0-1<0-1—u/C/0+0+1/D/feature_2
/E/content+3:3+4&0+3#2+0/F/content_1/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
t~u-l+a=f:1_1/A/0_0_2/B/0-0-1:3-1&7-5#3-3$3-3>1-0<1-0—no_vowels/C/1+1+1
/D/feature_2/E/content+3:3+4&0+3#2+0/F/content_1/G/6_2/H/11=6:4=1&L-L%/I/0_0
/J/23+15-4
u~l-a+f=o:1_1/A/0_0_1/B/1-1-1:1-1&8-4#3-2$3-2>2-0<2-0—a/C/1+1+4/D/content_3
/E/content+1:4+3&1+2#0+0/F/content_1/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
l~a-f+o=s:1_4/A/1_1_1/B/1-1-4:1-1&9-3#4-1$4-1>0-1<0-1—o/C/0+0+2/D/content_1

/E/content+1:5+2&2+1#0+0/F/content_2/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
a~f-o+s=t:2_3/A/1_1_1/B/1-1-4:1-1&9-3#4-1$4-1>0-1<0-1—o/C/0+0+2/D/content_1
/E/content+1:5+2&2+1#0+0/F/content_2/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
f~o-s+t=r:3_2/A/1_1_1/B/1-1-4:1-1&9-3#4-1$4-1>0-1<0-1—o/C/0+0+2/D/content_1
/E/content+1:5+2&2+1#0+0/F/content_2/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
o~s-t+r=@:4_1/A/1_1_1/B/1-1-4:1-1&9-3#4-1$4-1>0-1<0-1—o/C/0+0+2/D/content_1
/E/content+1:5+2&2+1#0+0/F/content_2/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
s~t-r+@=n:1_2/A/1_1_4/B/0-0-2:1-2&10-2#5-1$5-1>0-0<0-0—@/C/1+1+3/D/content_1
/E/content+2:6+1&3+0#0+0/F/feature_0/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
t~r-@+n=i:2_1/A/1_1_4/B/0-0-2:1-2&10-2#5-1$5-1>0-0<0-0—@/C/1+1+3/D/content_1
/E/content+2:6+1&3+0#0+0/F/feature_0/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
r~@-n+i=t:1_3/A/0_0_2/B/1-1-3:2-1&11-1#5-0$5-0>1-0<1-0—i/C/0+0+0/D/content_1
/E/content+2:6+1&3+0#0+0/F/feature_0/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
@~n-i+t=#:2_2/A/0_0_2/B/1-1-3:2-1&11-1#5-0$5-0>1-0<1-0—i/C/0+0+0/D/content_1
/E/content+2:6+1&3+0#0+0/F/feature_0/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
n~i-t+#=xx:3_1/A/0_0_2/B/1-1-3:2-1&11-1#5-0$5-0>1-0<1-0—i/C/0+0+0/D/content_1
/E/content+2:6+1&3+0#0+0/F/feature_0/G/6_2/H/11=6:4=1&L-L%/I/0_0/J/23+15-4
i~t-#+xx=xx:xx_xx/A/0_0_2/B/xx-xx-xx:xx-xx&xx-xx#xx-xx$xx-xx>xx-xx<xx-xx—xx/
C/0+0+0/D/content_1/E/xx+xx:xx+xx&xx+xx#xx+xx/F/feature_0/G/6_2/H/xx=xx:4=1
&L-L%/I/0_0/J/23+15-4

Appendix F

Sample List of Diphone Coverage Utterances

The list of the first 50 utterances in each subset is presented:

F.1 *diph1*

001. Nu este treaba lor ce constituţie avem.
002. Ea era tot timpul pe minge.
003. Nicoară crede că acest concurs va avea succes.
004. Afganistanul va fi reprezentat la adunarea generală de ministrul de externe, a declarat un responsabil al misiunii.
005. Evenimentul are ca scop facilitarea schimbului de idei privind viitorul securităţii energetice în aceste regiuni.
006. La serviciu vin dimineaţa, iar acasă ajung seara.
007. Am intervenit să aplanez conflictul.
008. Dacă lucrurile scapă de sub control.
009. Atenţie eu sunt şeful aici.
010. Domnul Bergodi era bine venit la meci chiar dacă dumnealui nu mai face parte din staf.
011. În caz contrar se va ajunge la amânarea prezentului.
012. Cum era să ratez meciul cu Rapid din cupa României.
013. Sud americanii sunt însă mai boemi.
014. Am hotărât să urmez cariera militară.
015. Sunt şi acele tatuaje o minciună.
016. Aici şi-a continuat studiile.
017. La acel restaurant avem asigurată masa.
018. Cei care l-au cunoscut au doar cuvinte de laudă despre fostul patron de la Irish Pub.
019. Nu e vina oamenilor că nu au de lucru.
020. În schimb va construi podul pe cheltuiala lui.
021. De multe ori jucam fotbal cu mingea de tenis.
022. Îmi plac foarte mult căţeii.
023. Aceasta va rămâne în libertate timp de trei luni pentru tratamente medicale.
024. Pe data de zece octombrie toate depozitele de gaze pe care le are România vor fi pline ochi.
025. Altfel cred că realităţile nu se vor schimba mult timp pe aici.

026. Îl vor redescoperi pe Marx.
027. Emisarul Statelor Unite Mark H.
028. Cei mai mulţi sunt zilieri.
029. Tirajul lor este simbolic, între două mii şi trei mii de exemplare zilnic.
030. Suspectul principal va fi reaudiat.
031. Semăn cu mamii şi cu tatii.
032. Gorjenii sunt liniştiţi vor evolua cu Oţelul.
033. Sunt resursele cele mai bogate comparativ cu restul resurselor care există, ceea ce ar putea însemna schimbarea industriei petroliere.
034. În opinia mea aici este cheia.
035. Aceasta nu aduce atingere rezultatului procedurii spun oficialii comisiei europene în comunicat.
036. E greu să coordonezi o echipă.
037. Nu va putea fi un meci echilibrat.
038. Sunteţi un cuplu de succes.
039. Un caz realmente impresionant l-am întâlnit la Iveşti.
040. Vrem cu orice preţ victoria, a spus portughezul.
041. Peste tot sunt locuri frumoase, femei frumoase.
042. În general se lucrează comenzi.
043. Vorbeşti despre Kovacs antrenorul ungur.
044. Nu ne-am considerat favoriţi danezii, au jucat modest dar numai pentru că noi am evoluat bine.
045. Cel mic nu dormea toată noaptea.
046. Aici a stat pentru Sfânta Liturghie.
047. Un personaj romantic foarte sensibil.
048. Nu au ce să ne ofere.
049. Traficul din Cluj e deranjant, nu-mi place.
050. Mi se pare extraordinară, a continuat el.

F.2 *diph2*

001. Volleiball club junior Delfin din Bucureşti organizează selecţie, pentru copii cu vârsta între şase şi treisprezece ani.
002. Acest lucru îl recunoaşte şi purtătorul de cuvânt al unităţii spitaliceşti, doctor Cristian Jianu.
003. Puneţi compoziţie în formă până la jumătate, montaţi în mijloc două ouă fierte.
004. Se serveşte cu sosul de muştar preparat din muştar diluat în apă şi oţet, sare piper şi ulei de măsline.
005. Min şaptezeci şi doi: Bujor ratează şansa de a înscrie dubla.
006. Fiecare clasă are dulăpioare pentru elevi calculatoare şi televizoare.
007. Am maşină am valoare.
008. A fost prima oară când i-am auzit vocea.
009. În caz contrar gorjenii riscau să nu mai joace duminică în campionat cu Oţelul.
010. Eugeniu Rădulescu tot aşa este.
011. Astăzi Andrei Pavel îl întâlneşte pe uruguayanul Pablo Cuevas.
012. Varianta cu cele mai mici şanse era la congres.
013. În paralel se injectează o substanţă care blochează accesul sangvin în locul respectiv.
014. Puneam versurile pe muzică afirmă digeiul.
015. Echipa Top Gear va filma si pe litoralul Mării Negre în Delta Dunării şi în podişul

Dobrogei.

016. Şi să iei măsuri.
017. Bucureştiul vieţuieşte între gri şi prăfuit.
018. Eu ce fac, dorm pe stradă.
019. Ceauşeasca crispată la faţă a luat scrisoarea.
020. Conducătorul auto Ilie g.
021. Am dosar la Hollywood.
022. Le ştiam deja după paşi.
023. Sau o schiţă.
024. Această cafea este servită în ceşcuţe foarte mici.
025. Lucrăm cu oameni săraci, iar când percepi o taxă se simt nedreptăţiţi.
026. Dan Cristea Veljovici şaizeci şi opt grozav Verdeş cincizeci şi opt.
027. Gorbaciov lui g.
028. Efectuăm evaluări la milioane chiar bilioane de bliţuri.
029. Pe se ve eindhoven.
030. Vom înfrunta o echipă solidă.
031. E Algeria condusă de Abdelaziz Bouteflika.
032. O lepră cultivată.
033. Tot ce nu clădeşte strică.
034. Alin însă parcă nici nu îi auzea.
035. Mihai i i ge.
036. Când trăiesc şi când mor.
037. Unu Europei pentru amatorii genului.
038. Anul acesta în urma României au stat naţiuni precum Austria, Germania sau Cehia.
039. Practicaţi zilnic exerciţiile kegel.
040. Solistă va fi pianista franceză Grimaud.
041. Se auzea prin mulţime.
042. Sau hai Kape pasă la Bănel.
043. Şaisprezece ianuarie.
044. În secolul nouăsprezece, patruzeci şi şase, se stabileşte definitiv la Puerto Rico.
045. E calomnie tot ce aşterneţi pe hârtie.
046. Nu poţi iubi ceva mai mult decât omul.
047. Vivu şi Petru Rareş.
048. Oameni insensibili, misogini, nenorociţi.
049. Bărbatul urmează a fi predat autorităţilor belgiene.
050. Şi totuşi o să -l facem.

Appendix G

Sample List of Random Utterances

The first 50 utterances in each subset are presented:

G.1 *rnd1*

001. De asemenea, contează şi dacă imobilul este la stradă sau nu.

002. Până în prezent, proiectul avea susţinerea ambelor partide, care şiau împărţit deja conducerea noilor entităţi.

003. Băimăreanul urăşte lipsa de punctualitate şi făţărnicia.

004. În acest cămin au prioritate studenţii în ani terminali.

005. Dincolo de efectele economice şi comerciale, greva ne face să avem şi lipsă de încredere în justiţie.

006. Dacă vecinii nu reclamă, noi nu putem depista spargerile de pereţi ilegale din apartamente.

007. Amploarea nu va fi aceeaşi, dar organizatorii se vor strădui să aducă produse autentic germane la poalele Tâmpei.

008. Vreau să continuu în acest domeniu până la sfârşitul vieţii spune cu o mină foarte serioasă micuţa.

009. Căzăturile sunt la ordinea zilei, am picioarele negre, sunt ca un dalmaţian.

010. Pe de altă parte, conform rezultatelor obţinute.

011. Nu este însă la fel de clar cu cât ar creşte consumul şi dependenţa de droguri.

012. Departajarea se face pe probe de cincizeci de metri care trebuiesc parcurse în zece secunde.

013. O păstrez ca amintire la loc de cinste, îşi aminteşte informaticianul.

014. Sabău spune că stilul de joc al lăncierilor este acelaşi ca şi pe vremea când juca el în Olanda.

015. M-am înţeles bine cu toată echipa, mi-a plăcut foarte mult spune ea.

016. Dacă îi prindem nu îi exmatriculăm din facultate.

017. Pe lângă câini bătrâni avem şi căţei foarte drăguţi şi inteligenţi.

018. Ion Tirinescu deţine funcţia de şef al poliţiei rutiere hunedoara din anul două mii cinci.

019. Răspunde nu numai de funcţionarea optimă ci şi de asigurarea banilor pentru toate cheltuielile radioului.

020. I-a plăcut jurnalismul şi pentru a ajunge să fie formator de opinie, a început facultatea la Piteşti.

021. Sunt unii care au descoperit pe aceste reţele foşti colegi de şcoală.

022. Arădenii sunt din ce în ce mai speriaţi la gândul că trebuie să circule pe drumul Arad

Şiria.

023. Pe toate rutele din judeţul Vâlcea se fac, începând de săptămâna aceasta, controale care vizează activităţile de transport persoane.

024. Capitala şi nord-estul au cunoscut cele mai mari ritmuri de creştere ale sumelor restante.

025. Vechile cabinete erau la o distanţă de o sută cincizeci metri de spital şi nu corespundeau reglementărilor ministerului sănătăţii.

026. Fără aceste două elemente de bază nu ai cum să lucrezi cu niciun copil.

027. Dorim să atragem finanţarea printrun program transfrontalier cu Republica Moldova.

028. La ora actuală oamenii sunt preocupaţi de a avea şi nu mai sunt preocupaţi de a fi.

029. Românii din străinătate nu există pentru autorităţile române de acolo a oftat mama bărbatului.

030. De asemenea este organizat şi campionatul naţional de car-audio.

031. Printre taxele cele mai mari le are facultatea de drept.

032. Comitetul de implementare are rolul de a supraveghea îndeplinirea obligaţiilor ce revin părţilor la convenţie.

033. Locaţia a fost stabilită de către inspectoratul şcolar Arad la grădiniţa cu program prelungit, situată în centrul oraşului.

034. Precizează Gheorghe Crivac, unul dintre prorectorii universităţii Piteşti.

035. Uneori suntem chiar mai bune decât ei.

036. Compania mai are activităţi în domenii precum imobiliare, tehnologie şi energie.

037. Am fost cu mama şi cu sora mea.

038. Din prima până în ultima secundă, convorbirea telefonică e o beşteleală de cartier aplicată cu toată impetuozitatea sărmanului cetăţean.

039. În localitate există şase poliţişti, dintre care unul este în concediu şi unul are mâna ruptă.

040. Încă de la prima accesare m-a atras foarte mult.

041. Cât durează până câştigă, nu poate spune.

042. Dacă şi pe data de treizeci septembrie se întârzie cu primirea avansului, pe doi octombrie la ora şapte.

043. Copiii care ţineau în braţe bărcuţe cu telecomandă au înconjurat fântâna speranţei, gata de cursă.

044. În blocul şase, două scări, adică aproximativ treizeci de apartamente, erau terminate când au venit actele de la primărie.

045. A fost testat alcooltest, rezultatul fiind negativ.

046. Până la urmă rămâi oricum doar cu amintirile plăcute, oboseala trece cu un somn bun.

047. De atunci, am învăat foarte multe persoane să danseze pe diferite ritmuri.

048. După ce i-au administrat îngrijirile de urgenă, medicii l-au dus la reanimare, unde se află sub supraveghere de specialitate.

049. Viorica Crişan, directorul muzeului, stă zilnic peste program la serviciu.

050. Eşecul statului federal în faţa cicloanelor rămâne unul din cele mai profunde stigmate ale preşedinţiei lui George.

G.2 *rnd2*

001. Ne doare pe toţi, încă de a doua zi dimineaţă.

002. Drumul a fost blocat mai multe ore.

003. Acum am fost penalizaţi şi cu penalty.

004. Această partidă se poate compara cu meciul de la Doneţk.

005. Gâtul a necesitat intervenţie de lungă durată.

006. Îmi dezvoltă o anumită minuţiozitate.
007. Îşi încheie comentariul Ria Novosti.
008. Dacă nu s-a intervenit la timp, normal că acum sunt grămezi uriaşe.
009. Pot să spun că mi-a schimbat viaţa.
010. Primul meu antrenor de tenis a fost profesorul Geantă.
011. Apoi au venit foarte multe delegări în toată ţara.
012. Am avut însă doar doi ani deschisă o casă de modă.
013. În două mii unu a obţinut titlul de master series, la Montreal.
014. Aş vrea să se califice amândouă.
015. Staţia de la Penny Market nu va fi funcţională pe durata lucrărilor.
016. El a reuşit să-l imobilizeze pe bărbat.
017. Preţul unei ore de joacă sub supravegherea personalului angajat costă şapte lei.
018. Suntem prieteni cu skaterii.
019. De altfel, am vrut să mă şi retrag.
020. A fost foarte greu pentru ea în ultima vreme.
021. Atât autotrenul cât şi cantitatea de ţigări au fost predate inspectorilori vamali.
022. Oncescu s-a supărat pe arbitrii care n-au fost atenţi la meciurile lui.
023. Mai mulţi agenţi sub acoperire filmează cu o cameră ascunsă, toate neregulile.
024. Avem ocazia să prezentăm un front unit.
025. Sunt doar din ce în ce mai mulţi.
026. Reţeaua Al Qaida a revendicat atacurile.
027. În primele zile nu facem nici douăzeci lei, arată Rozalia.
028. Kone a şutat din lovitură liberă peste poarta giuleşteană.
029. Cronjaeger a renunţat să ceară o contraexpertiză.
030. Familia Lăcătuş ş-ia făcut un trecut în domeniul vânzării castanelor.
031. Apă otrăvită, după cum veţi vedea în cele ce urmează.
032. Eu nu mai înţeleg pentru ce dau acei bani.
033. Îmi place mult să călătoresc, să văd lucruri noi.
034. E un sentiment aparte când îi vezi pe toţi că dansează.
035. Aceeaşi situaţie s-a înregistrat şi în judeţul Vaslui, la vama Albiţa.
036. Au trecut mulţi ani de când am lucrat împreună.
037. După o vară plină de tensiune, învăţământul românesc trece într-o altă etapă.
038. Am jucat bine, dar nu vreau să mă mai avânt.
039. Răbdarea este o calitate pe care nu o deţin mulţi dascăli.
040. România aceea e la sat.
041. Înainte, am fost fierar betonist la primărie, povesteşte bărbatul.
042. Două dintre echipele româneşti au avut reproşuri la adresa arbitrajelor.
043. Irlanda respinge prin referendum ratificarea tratatului.
044. În acest caz, nu se poate spune de o schimbare.
045. Din nefericire, sunt unele persoane din sectorul financiar care înţeleg greşit momentul.
046. Artistului nu-i place neseriozitatea persoanelor cu care lucrează.
047. Aşa că am reuşit să strâng mai multe bucăţi din acelaşi manual.
048. Declară viceprimarul Cornel Ionică.
049. Cred că este nevoie doar de puţină încredere.
050. Asociaţia nu avea statut legal atunci când am devenit eu preşedinte.

G.3 *rnd3*

001. Am crezut că femeia din faţa mea a murit.

002. El crede că dublarea nu este neapărat un dezavantaj pentru italieni.
003. Pe ce personaje mizaţi în noul sezon, toate personajele sunt o mizã.
004. Cu toţii ne aducem aminte de aceste momente.
005. Reprezentantul sindicatului medicilor specialişti din Botoşani a reacţionat dur.
006. Astfel se răspândeşte ghinionul.
007. Ne plătim datoriile adunate în vară.
008. Aceasta este o idee bună.
009. La faţa locului s-au deplasat mai multe echipaje de prim ajutor.
010. Talk show-ul s-a transformat dintro dezbatere în spectacol.
011. Totuşi, Anderlecht reprezintă una dintre principalele pretendente la calificare în Europa League.
012. Covrigii fierţi sunt la mare căutare în rândurile acestora.
013. Potrivit organizatorilor, bugetul festivalului se ridică la două sute mii de euro.
014. Numărătoarea inversă a început.
015. Dacă fura, vă daţi seama că nu îmi mai cerea mie bani.
016. Nu se face aşa ceva, a spus Istudor.
017. Spune tânăra cântăreaţă.
018. Oamenii se simţeau legaţi de acel loc.
019. Ba din contră, m-au susţinut, mai spune adolescenta.
020. Îmi place foarte mult să merg la raliuri.
021. Ofiţerii din cadrul I J P F.
022. Acesta putea afecta circuitele electronice ale micilor ambarcaţiuni.
023. Iniţial, lucrările au început în vara anului trecut.
024. Atunci nu aveam timp de vizite, jucam şi dormeam.
025. Avem tot ce ne trebuie şi nici nu ne costă mult.
026. Dacă era nouă zero, nu era nicio problemă, nu comenta nimeni.
027. Nu vă mai complicaţi viaţa cu Plevuşcă.
028. Arunc boabele crăpate şi nu le păstrez decât pe cele sănătoase.
029. Steaua este o echipă de calitate, cu mare tradiţie în cupele europene.
030. Îmi plac şi micile şmecherii afirmă arădeanca.
031. El reprezintă districtul treisprezece, care include cartierul Queens.
032. Într-un cuvânt, Rusia dintotdeauna.
033. Finalul a fost dramatic pentru scandinav.
034. La rândul său, a oferit publicului treizeci de mingi cu autograful lui.
035. Tânăra are permisul de conducere de doar un an şi jumătate.
036. Imediat au anunţat poliţia de la secţia trei.
037. Ocupantele locurilor doi - patru vor merge în cupa EHF.
038. Poliţia comunală Mătăsari a fost sesizată aseară de către F punct V.
039. În schimb, firmele lui Penescu primeau doar contravenţii.
040. La fel, avizul de gospodărire a apelor este valabil doar cinsprezece zile.
041. La fel şi cele ale copiilor, ingenue şi pline de afecţiune.
042. Preţul scade cu încă un pas, la o sută patru mii euro.
043. Închiderea acestui buget, la cifrele convenite, va fi oricum o mare performanţă.
044. Nu îmi place singurătatea şi de aceea iubesc metropolele, spune scenarista.
045. Szekely, douăzeci şi cinci.
046. Mi-a zis cineva de la CFR că Bergodi e cam tralala.
047. Statul nu trebuie să mai fie o povară pentru cetăţeni.
048. Pentru evoluţiile lui Onicaş şi Ionescu, singura explicaţie poate fi oboseala.
049. El nu a explicat de ce Karzai a renunţat la deplasare.
050. Este vorba despre AGD.

Appendix H

Sample List of Fairytale Utterances

The first 50 utterances according to the performed segmentation are presented:

H.1 *Povestea lui Stan Păţitul*

001. Era odată un flăcău stătut, pe care-l chema Stan.

002. Şi flăcăul acela din copilăria lui se trezise prin străini, fără să cunoască tată şi mamă şi fără nici o rudă care să-l ocrotească şi să-l ajute.

003. Şi, ca băiat străin ce se găsea, nemernicind el de colo până colo pe la uşile oamenilor, de unde până unde s-a oploşit de la o vreme într-un sat mare şi frumos.

004. Şi aici, slujind cu credinţă ba la unul, ba la altul, până la vârsta de treizeci şi mai bine de ani, şi-a sclipuit puţine parale, câteva oi, un car cu boi şi o văcuşoară cu lapte.

005. Mai pe urmă şi-a înjghebat şi o căsuţă, şi apoi s-a statornicit în satul acela pentru totdeauna, trăgânduse la casa lui şi muncind ca pentru dânsul.

006. Vorba ceea. Şi piatra prinde muşchi dacă şede mult întrun loc.

007. Şi cum s-a văzut flăcăul cu casă şi avere bunicică, nu mai sta locului, cum nu stă apa pe pietre, şi mai nu-l prindea somnul de harnic ce era.

008. Dintro parte venea cu carul, în alta se ducea, şi toate treburile şi le punea la cale singurel.

009. Nu-i vorbă că, de greu, greu îi era. Pentru că, în lipsa lui, n-avea cine săi -îngrijească de casă şi de vitişoare cum trebuie.

010. Numai, dă, ce să facă bietul om. Cum era să se întindă mai mult, că de-abia acum se prinsese şi el cu mâinile de vatră. Şi câte a tras până s-a văzut la casa lui, numai unul Dumnezeu ştie.

011. De aceea alerga singur zi şi noapte în toate părţile, cum putea, şi muncea în dreapta şi în stânga, că doar doar a încăleca pe nevoie, şapoi atunci, văzând şi făcând.

012. Toate ca toate, dar urâtul îi venea de hac.

013. În zile de lucru, calea valea. se lua cu treaba şi uita de urât.

014. Dar în nopţile cele mari, când era câte o givorniţă cumplită şi se mai întâmpla să fie şi sărbători la mijloc, nu mai ştia ce să facă şi încotro să apuce, vorba cântecului.

015. De urât mă duc de acasă.

016. Şi urâtul nu mă lasă.

017. De urât să fug în lume.

018. Urâtul fuge cu mine.

019. Se vede lucru că aşa e făcut omul, să nu fie singur.

020. De multe ori i-a venit flăcăului în cap să se însoare, dar când îşi aducea aminte uneori

de câte i-au spus că au pătimit unii şi alţii de la femeile lor, se lua pe gânduri şi amâna, din zi în zi şi de joi până mai de apoi, această poznaşă trebuşoară şi gingaşă în multe privinţe, după cum o numea el, gândindu-se mereu la multe de toate.

021. Unii zic aşa, că femeia-i sac fără fund.

022. Cea mai fi şi asta. Alţii, că să te ferească Dumnezeu de femeia leneşă, mârşavă şi risipitoare. alţii alte năstruşnicii, încât nu ştii ce să crezi şi ce să nu crezi.

023. Numai nu-i vorbă că am văzut eu şi destui bărbaţi mult mai ticăiţi şi mai chitcăiţi decât cea mai bicisnică femeie.

024. Şi aşa, trezindu-se el în multe rânduri vorbind singur, ca nebunii, sta în cumpene, să se însoare.

025. să nu se însoare.

026. Şi, ba s-a însura la toamnă, ba la iarnă, ba la primăvară, ba la vară, ba iar la toamnă, ba vremea trece, flăcăul începe şi el a se trece, mergând tot înainte cu burlăcia, şi însurătoarea rămâne baltă.

027. Şi apoi este o vorbă: că până la douăzeci de ani se însoară cineva singur, de la douăzeci la douăzeci şi cinci îl însoară alţii. de la douăzeci şi cinci la treizeci îl însoară o babă, iară de la treizeci de ani înainte numai dracu îi vine de hac.

028. Tocmai aşa s-a întâmplat şi cu flăcăul acesta că, până la vremea asta, nici el de la sine, nici prietenii, nici babele câtu-s ele dea dracului, de prefăcute şi iscoditoare tot nu lau putut face să se însoare.

029. Stan era om tăcut în felul său, dar şi când da câteo vorbă dintr-însul vorba era vorbă, la locul ei, şi nu-l putea răpune te miri cine.

030. Mulţi trăgeau nădejdea să-l ia de ginere, dar flăcăul era chitit la capul său şi nu se da cu una, cu două.

031. Şi aşa, de la o vreme, şi prietenii şi babele, lehămetindu-se, l-au dat în burduful dracului şi l-au lăsat pe seama lui, să facă de acum înainte ce-a şti el cu dânsul, că ei şi-au luat toată nădejdea.

032. Amu, întruna din zile, flăcăul se scoală de noapte, face mămăligă îmbrânzită şi cea mai dat Dumnezeu, pune mâncarea în traistă, înjugă boii la car, zice Doamne ajută şi se duce la pădure, să-şi aducă un car de lemne.

033. Şi ajungând el în pădure pe când se mijea de ziuă, a tăiat lemne, a încărcat carul zdravăn şi l-a cetluit bine, şi pân-or mai mânca boii, s-a pus să mănânce şi el ceva.

034. Şi după ce a mâncat cât a trebuit, i-a mai rămas o bucăţică de mămăligă îmbrânzită şi, făcând o boţ, a zis. Ce s-o mai duc acasă. ia so pun ici pe teşitura asta, că poate a găsi o vreo lighioaie ceva, a mânca o şi ea ş-a zice o bodaproste.

035. Şi punând mămăliga pe teşitură, înjugă boii, zice iar un Doamne ajută şi, pe la prânzişor, porneşte spre casă.

036. Şi cum a pornit el din pădure, pe loc s-a şi stârnit un vifor cumplit, cu lapoviţă în două, de nu vedeai nici înainte, nici înapoi.

037. Mânia lui Dumnezeu ce era afară. să nu scoţi câine din casă, dar încă om. însă dracul nu caută mai bine. La aşa vreme te face să pierzi răbdarea şi, fără să vrei, te vâră în păcat.

038. În acea zi, Scaraoschi, căpetenia dracilor, voind a-şi face mendrele cum ştie el, a dat poruncă tuturor slugilor sale ca să apuce care încotro a vedea cu ochii, şi pretutindene, pe mare şi pe uscat, să vâre vrajbă între oameni şi să le facă pacoste.

039. Atunci dracii s-au împrăştiat, iute ca fulgerul, în toate părţile.

040. Unul din ei a apucat spre păduri, să vadă de na putea trebălui ceva şi pe acolo. doar a face pe vrun om să bârfească împotriva lui Dumnezeu, pe altul să-şi chinuiască boii, altuia să-i rupă vrun capăt sau altceva de la car, altuia să-i schilodească vrun bou, pe alţii săi facă să se bată până s-or ucide, şi câte alte bazaconii şi năzbutii de care iscodeşte şi vrăjeşte dracul.

041. Ce-or fi isprăvit ceilalţi draci nu ştim, dar acestui de la pădure nu i-a mers în acea zi.

042. s-a pus el, nu-i vorbă, luntre şi punte ca să-şi vâre codiţa cea bârligată undeva, dar degeaba i-a fost, că, pe unde se ducea, tot în gol umbla.

043. Şi tot cercând el ba ici, ba colea, înspre seară numai ce dă de-o pârtie.

044. Atunci se ia tiptil tiptil pe urma ei şi se duce tocmai la locul de unde încărcase Stan lemnele.

045. Şi, când colo, găseşte numai locul, pentru că flăcăul, după cum am spus, de mult ieşise din pădure şi se dusese în treaba lui.

046. Văzând el dracul că nici aici n-a izbutit nimica, crâşcă din măsele şi crapă de ciudă, pentru că era îngrijit cu ce obraz să se înfăţişeze înaintea lui Scaraoschi. Şapoi, afară de aceasta, era buimac de cap şi hămesit de foame, de atâta umblet.

047. Şi cum sta el pe gânduri, posomorât şi bezmetic, numai iaca ce vede pe-o teşitură un boţ de mămăligă.

048. Atunci, bucuria dracului. Odată o şi haleşte şi nu zice nimica.

049. Apoi, nemaiavând ce face, îşi ia coada între vine şi se întoarce la stăpânu său, şi, cum ajunge în iad, Scaraoschi îl întreabă.

050. Ei, copile, ce ispravă ai făcut. Câte suflete mi-ai arvonit. Dă-ţi solia.

H.2 *Ivan Turbincă*

001. Era odată un rus, pe care îl chema Ivan.

002. Şi rusul acela din copilărie se trezise în oaste.

003. Şi slujind el câteva soroace de-a rândul, acuma era bătrân.

004. Şi maimarii lui, văzându-l că şi-a făcut datoria de ostaş, l-au slobozit din oaste, cu arme cu tot, să se ducă undea vrea, dându-i şi două carboave de cheltuială.

005. Ivan atunci mulţumi maimarilor săi şi apoi, luându-şi rămas bun de la tovarăşii lui de oaste, cu care mai trase câte-o duşcă, două de rachiu, porneşte la drum cântând.

006. Şi cum mergea Ivan, şovăind când la o margine de drum, când la alta, fără să ştie unde se duce, puţin mai înaintea lui mergeau din întâmplare, pe-o cărare lăuntrică, Dumnezeu şi cu Sfântul Petre, vorbind ei ştiu ce.

007. Sfântul Petre, auzind pe cineva cântând din urmă, se uită înapoi şi, când colo, vede un ostaş mătăhăind pe drum în toate părţile.

008. Doamne, zise atunci Sfântul Petre, speriat. ori hai să ne grăbim, ori să ne dăm într-o parte, nu cumva ostaşul cela să aibă harţag, şi să ne găsim beleaua cu dânsul.

009. Ştii c-am mai mâncat eu o dată de la unul ca acesta o chelfăneală.

010. N-ai grijă, Petre, zise Dumnezeu.

011. De drumeţul care cântă să nu te temi.

012. Ostaşul acesta e un om bun la inimă şi milostiv.

013. Vezil. Are numai două carboave la sufletul său. Şi, drept cercare, hai, făte tu cerşetor la capătul ist de pod, şi eu la celălalt.

014. Şi să vezi cum are să ne dea amândouă carboavele de pomană, bietul om. Aduţi aminte, Petre, de câte ori ţi-am spus, că unii ca aceştia au să moştenească împărăţia cerurilor.

015. Atunci Sfântul Petre se pune jos la un capăt de pod, iară Dumnezeu la celălalt şi încep a cere de pomană.

016. Ivan, cum ajunge în dreptul podului, scoate cele două carboave de unde le avea strânse şi dă una lui Sfântul Petre şi una lui Dumnezeu, zicând.

017. Dar din dar se face raiul.

018. Na-vă. Dumnezeu mia dat, eu dau, şi Dumnezeu iar mi-a da, că are de unde.

019. Şi apoi Ivan începe iar a cânta şi se tot duce înainte.

020. Atunci Sfântul Petre zice cu mirare.

021. Doamne, cu adevărat bun suflet de om e acesta, şi n-ar trebui să meargă nerăsplătit de la faţa ta.

022. Dar, Petre, las că am eu purtare de grijă pentru dânsul.

023. Apoi Dumnezeu porneşte cu Sfântul Petre şi, cât ici, cât cole, ajung pe Ivan, care-o ducea tot întrun cântec, de parcă era toată lumea a lui.

024. Bună calea, Ivane, zise Dumnezeu.

025. Dar cânţi, cânţi, nu te-ncurci.

026. Mulţumesc dumneavoastră, zise Ivan, tresărind.

027. Dar de unde ştii aşa de bine că mă cheamă Ivan.

028. Dapoi, dacă noi şti eu, cine altul are să ştie. răspunse Dumnezeu.

029. Dar cine eşti tu, zise Ivan cam zborşit, de te lauzi că ştii toate.

030. Eu sunt cerşetorul pe care l-ai miluit colo la pod, Ivane.

031. Şi cine dă săracilor împrumută pe Dumnezeu, zice scriptura.

032. Na-ţi împrumutul înapoi, căci noi nu avem trebuinţă de bani.

033. Ia, numai am vrut să dovedesc lui Petre cât eşti tu de milostiv.

034. Află acum, Ivane, că eu sunt Dumnezeu şi pot să-ţi dau orice-i cere de la mine. pentru că şi tu eşti om cu dreptate şi darnic.

035. Ivan atunci, cuprins de fiori, pe loc s-a dezmeţit, a căzut în genunchi dinaintea lui Dumnezeu şi a zis.

036. Doamne, dacă eşti tu cu adevărat Dumnezeu, cum zici, rogute blagosloveştemi turbinca asta, ca ori pe cine-oi vrea eu, să-l vâr într-însa. Şi apoi să nu poată ieşi de aici fără învoirea mea.

037. Dumnezeu atunci, zâmbind, blagoslovi turbinca, după dorinţa lui Ivan, şi apoi zise.

038. Ivane, când te-i sătura tu de umblat prin lume, atunci să vii să slujeşti şi la poarta mea, căci nu ţ-ia fi rău.

039. Cu toată bucuria, Doamne. am să vin numai decât, zise Ivan.

040. Dar acum, deodată, mă duc să văd, nu mi-a pica ceva la turbincă.

041. Şi zicând aceste, apucă peste câmpii de-a dreptul, spre nişte curţi mari, care deabia se zăreau înaintea lui, pe culmea unui deal.

042. Şi merge Ivan, şi merge, şi merge, până când, pe înserate, ajunge la curţile cele.

043. Şi cum ajunge, intră în ogradă, se înfăţişează înaintea boierului şi cere găzduire.

044. Boierul acela cică era cam zgârcit, dar, văzând că Ivan este om împărătesc, n-are ce să facă.

045. Şi vrând nevrând, porunceşte unei slugi să dea lui Ivan ceva de mâncare şi apoi să-l culce în nişte case nelocuite, unde culca pe toţi musafirii care veneau aşa, nitam nisam.

046. Sluga, ascultând porunca stăpânului, ia pe Ivan, îi dă ceva de mâncare şi apoi îl duce la locul hotărât, să se culce.

047. Las dacă nu i-a da odihna pe nas, zise boierul în gândul său, după ce orândui cele de cuviinţă.

048. Ştiu că are să aibă de lucru la noapte.

049. Acum să vedem care pe care. Ori el pe draci, ori dracii pe dânsul.

050. Căci trebuie să vă spun că boierul acela avea o pereche de case, mai de-o parte, în care se zice că locuia necuratul.

Appendix I

List of Semantically Unpredictable Sentences for Romanian

001. Semnul glumeşte din capul major.
002. Cardul curge la lista viitoare.
003. Ardeiul gol laudă oceanul.
004. Vulpea temută zice sacul.
005. Pierde paiul sau poporul.
006. Cât prescrie rucsacul bobul acru?
007. Cum lipeşte căminul ciclul bun?
008. Cortul descurcă şahul care usucă.
009. Verbul iese fără căţelul rău.
010. Numele fuge spre gamba uzată.
011. Lupul timid refuză albumul.
012. Epoca demnă exportă zidul.
013. Crează lemnul şi dragonul.
014. Unde invită aerul maistrul cinic?
015. Când aleargă stimulul pomul solar?
016. Imnul găseşte neamul care decide.
017. Şamponul doarme după butonul sec.
018. Cablul leşină sub cireaşa groasă.
019. Uleiul uzat convinge paharul.
020. Morala amară lasă afişul.
021. Ţine capacul sau deceniul.
022. Cum duce şirul secolul şocat?
023. Unde transportă inelul ciorapul lacom?
024. Lucrul percepe genul care deschide.
025. Garajul tremură în aburul pur.
026. Borcanul vine despre toamna verde.
027. Primarul alb agită şanţul.
028. Vila bună probează scaunul.
029. Manevrează gerul şi ţipătul.
030. Cât scoate regimul ficatul atent?

031. Când editează saltul fanul tocit?
032. Balonul roade podul care depune.
033. Procentul constată peste anul ud.
034. Pragul rezistă cu apa neagră.
035. Becul dur extinde zâmbetul.
036. Ruda fină salută bonusul.
037. Testează polenul şi pionul.
038. Unde invită polenul căpitanul enorm?
039. Cât leagă ambalajul banul urban?
040. Vaporul ascultă duşul care reţine.
041. Gardul alunecă lângă nucul des.
042. Scutul survine sub vara sătulă.
043. Actorul tuns sare semnul.
044. Poşeta majoră editează spiritul.
045. Lasă camionul sau conul.
046. Cum alege contul dintele fals?
047. Când opreşte ţărmul bradul brun?
048. Cerul scrie visul care ţine.
049. Cuţitul pare lângă tigrul moale.
050. Darul uită spre găina utilă.
051. Puştiul fix remarcă trenul.
052. Reţeaua rece pozează aroma.
053. Acceptă ciocanul sau farul.
054. Cât ceartă decorul zmeul liber?
055. Când strică versul timpul mic?
056. Lacul gustă tortul care imită.
057. Oţelul tresare cu eroul spart.
058. Tânărul râde după aţa largă.
059. Peştele fin reface fulgerul.
060. Vremea toxică explorează cazanul.
061. Taie ţapul şi cheful.
062. Cum scoate bancul ursul adult?
063. Unde felicită şeful doctorul mut?
064. Tunetul egalează bobul care refuză.
065. Preţul cântă peste şarpele lung.
066. Templul alege despre taxa frumoasă.
067. Miezul ars ignoră semnalul.
068. Pudra lată îmbină nodul.
069. Exclude sportul şi volanul.
070. Cum uneşte jocul cărbunele dator?
071. Când crapă textul norul real?
072. Ţinutul include vagonul care crede.
073. Racul luptă din pisoiul drag.
074. Adevărul vibrează fără baia obosită.
075. Omul des maschează izvorul.

076. Steaua liberă aprobă gazul.
077. Iubeşte nisipul sau cabinetul.
078. Unde fereşte actul ochiul lacom?
079. Cât calcă schimbul soţul cinic?
080. Dansul cară palatul care pictează.
081. Etajul pică în iepurele ud.
082. Zarul fumegă la culmea blondă.
083. Atomul pur opreşte aparatul.
084. Caseta vastă preia avionul.
085. Redă arcul sau iazul.
086. Cum încearcă votul papucul fals?
087. Cât propune ciobul melcul dator?
088. Cercul compune oraşul care sapă.
089. Satul tuşeşte despre timpul fix.
090. Altarul tremură spre secţia secată.
091. Pictorul drag înscrie târgul.
092. Bursa grasă denunţă desenul.
093. Alege orezul şi minutul.
094. Unde ştie salonul colţul gol?
095. Când prinde hanul judeţul şocat?
096. Stratul aude filmul care cheamă.
097. Cadrul cade peste beciul atent.
098. Fânul zâmbeşte în camera rumenă.
099. Băiatul brun adaugă avansul.
100. Vocea rară asistă şocul.
101. Asistă metalul sau satul.
102. Cât judecă eseul părintele adult?
103. Când trimite ghiveciul soarele mut?
104. Osul scufundă testul care susţine.
105. Aurul zace la băiatul mic.
106. Metalul merge din făina rurală.
107. Vulcanul urban vede farul.
108. Soba iute apucă focul.
109. Înfundă terenul şi patul.
110. Cum citeşte cojocul ţăranul acru?
111. Unde rupe lacătul argintul enorm?
112. Fierul ridică salamul care filmează.
113. Scrumul renunţă lângă fratele liber.
114. Obiectul compară cu barca sărată.
115. Tonul dur cere cuibul.
116. Raza crudă reduce spaţiul.
117. Stinge culoarea şi ecranul.
118. Cât implică stocul pilotul rău?
119. Unde bea serialul moşul tuns?
120. Juriul sprijină procesul care aplaudă.

121. Creionul intră după bunicul fin.
122. Anunţul revine fără poza stoarsă.
123. Berbecul tocit foloseşte plumbul.
124. Limba uscată dezgheaţă chipul.
125. Dă mesajul sau regatul.
126. Cum arde cimentul vărul uzat?
127. Când ajută iaurtul sacul spart?
128. Teatrul lansează parcul care reclamă.
129. Tractorul suspină sub cerbul solar.
130. Unghiul stă peste sora strâmbă.
131. Piciorul alb deţine digul.
132. Masa mică pomeneşte corpul.
133. Spală steagul sau geamul.
134. Când coace lanţul calul timid?
135. Cât povesteşte rândul tonul lung?
136. Turnul închide ziarul care enumeră.
137. Basmul iese fără regele bun.
138. Anunţul umblă în cenuşa săracă.
139. Carul major roagă cadoul.
140. Uşa rotundă are discul.
141. Apelează liceul şi castelul.
142. Cum arată tariful fiul moale?
143. Unde răceşte malul căţelul real?
144. Excesul ia testul care sună.
145. Norocul oftează lângă atomul ars.
146. Efectul dispare despre fraza caldă.
147. Soarele sec preface paharul.
148. Hoaţa murdară ajunge hotelul.
149. Îngraşă vârful şi costumul.
150. Unde numără fardul ciorapul brun?
151. Cum zideşte grupul puştiul tocit?
152. Mărul acuză bolul care aduce.
153. Miezul tace cu coţLul moale.
154. Simbolul fuge la faţa străină.
155. Miezul liber suceşte secolul.
156. Doica fină distinge câmpul.
157. Vrea uleiul sau accentul.
158. Cât verifică untul bradul enorm?
159. Când ocupă catalogul doctorul major?
160. Astrul întreabă acul care agaţă.
161. Apusul stă după căpitanul rău.
162. Nivelul intră sub banca toxică.
163. Cerbul timid ascultă valul.
164. Clipa liberă trimite degetul.
165. Usucă cheful sau podul.

166. Unde reţine duşul fiul cinic?
167. Când explorează semnul tonul solar?
168. Zarul ia tânărul care bea.
169. Cabinetul rezistă spre maistrul alb.
170. Procesul tremură din roata rotundă.
171. Moşul acru povesteşte terenul.
172. Cheia demnă iubeşte hanul.
173. Agaţă patul şi metalul.
174. Cum îmbină borcanul nucul gol?
175. Cât vede anunţul melcul mut?
176. Inelul cară paiul care opreşte.
177. Secolul iese spre bunicul tuns.
178. Spiritul renunţă fără ploaia vastă.
179. Ursul real calcă plumbul.
180. Şina rece opreşte lemnul.
181. Crede gerul şi ghiveciul.
182. Unde găseşte eseul banul lung?
183. Cum ţine acul pictorul şocat?
184. Fânul are aerul care preia.
185. Balonul survine după calul lacom.
186. Testul vibrează la fata frumoasă.
187. Ciclul fin ridică lucrul.
188. Poşta murdară ceartă testul.
189. Crează nivelul sau satul.
190. Cât filmează şeful beciul sec?
191. Când decide votul sacul dur?
192. Hotelul refuză dansul care alege.
193. Ţapul glumeşte cu ţăranul des.
194. Scaunul dispare peste figura temută.
195. Pilotul mic roagă ciocanul.
196. Cina amară editează parcul.
197. Încearcă tariful şi basmul.
198. Cât acceptă discul butonul atent?
199. Unde lasă astrul aburul bun?
200. Aparatul alege castelul care pomeneşte.
201. Degetul cade în primarul ars.
202. Osul zâmbeşte din coaja caldă.
203. Peştele fals asistă contul.
204. Gena sărată prescrie semnalul.
205. Scrie adevărul sau bolul.
206. Când depune ciobul părintele pur?
207. Cum zice spaţiul vulcanul urban?
208. Simbolul fereşte cazanul care remarcă.
209. Aurul râde despre berbecul dator.
210. Avansul leşină lângă sarea lată.

211. Papucul ud redă lacul.
212. Marca rară agită şahul.
213. Exportă saltul sau ziarul.
214. Unde probează obiectul vărul drag?
215. Când arată farul dintele fix?
216. Efectul invită rucsacul care aprobă.

Appendix J

Selected Published Papers

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 442–450
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate

Adriana Stan [b,*], Junichi Yamagishi [a], Simon King [a], Matthew Aylett [c]

[a] The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, UK
[b] Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., 400027 Cluj-Napoca, Romania
[c] CereProc Ltd., Appleton Tower, 11 Crichton Street, Edinburgh EH8 9LE, UK

Received 13 July 2010; received in revised form 6 December 2010; accepted 6 December 2010
Available online 17 December 2010

Abstract

This paper first introduces a newly-recorded high quality Romanian speech corpus designed for speech synthesis, called "RSS", along with Romanian front-end text processing modules and HMM-based synthetic voices built from the corpus. All of these are now freely available for academic use in order to promote Romanian speech technology research. The RSS corpus comprises 3500 training sentences and 500 test sentences uttered by a female speaker and was recorded using multiple microphones at 96 kHz sampling frequency in a hemianechoic chamber. The details of the new Romanian text processor we have developed are also given.

Using the database, we then revisit some basic configuration choices of speech synthesis, such as waveform sampling frequency and auditory frequency warping scale, with the aim of improving speaker similarity, which is an acknowledged weakness of current HMM-based speech synthesisers. As we demonstrate using perceptual tests, these configuration choices can make substantial differences to the quality of the synthetic speech. Contrary to common practice in automatic speech recognition, higher waveform sampling frequencies can offer enhanced feature extraction and improved speaker similarity for HMM-based speech synthesis.

Keywords: Speech synthesis; HTS; Romanian; HMMs; Sampling frequency; Auditory scale

1. Introduction

Romanian is an Indo-European Romance language and has similarities with Italian, French and Spanish. Due to foreign occupation and population migration through the course of history, influences of various languages such as those of the Slavic family, Greek and Hungarian can be found in the Romanian language.

Currently, there are very few Romanian text-to-speech (TTS) systems: Most systems are still based on diphones (Ferencz, 1997) and the quality is relatively poor. To the best of our knowledge, only Ivona provides commercially-acceptable good quality Romanian synthesis; it is based on unit selection (Black and Cambpbell, 1995; Hunt and Black, 1996).[1] For promoting Romanian speech technology research, especially in speech synthesis, it is therefore essential to improve the available infrastructure, including free large-scale speech databases and text-processing front-end modules.

With this goal in mind, we first introduce a newly recorded high-quality Romanian speech corpus called "RSS",[2] then we describe our Romanian front-end modules and the speech synthesis voices we have built.

* Corresponding author.
E-mail addresses: adriana.stan@com.utcluj.ro (A. Stan), jyamagis@staffmail.ed.ac.uk (J. Yamagishi), simon.king@ed.ac.uk (S. King), matthew@cereproc.com (M. Aylett).

[1] See respectively http://tcts.fpms.ac.be/synthesis/mbrola.html, http://www.baum.ro/index.php?language=ro&pagina=ttsonline, and http://www.ivona.com for Romanian diphone system provided by the MBROLA project, Baum Engineering TTS system, Ancutza, and Ivona unit selection system.
[2] Available at http://octopus.utcluj.ro:56337/RORelease/.

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.12.002

HMM-based statistical parametric speech synthesis (Zen et al., 2009) has been widely studied and has now become a mainstream method for text-to-speech. The HMM-based speech synthesis system HTS (Zen et al., 2007c) is the principal framework that enables application of this method to new languages; we used it to develop these Romanian voices. It has the ability to generate natural-sounding synthetic speech and, in recent years, some HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems (Karaiskos et al., 2008) in terms of naturalness and intelligibility. However, relatively poor perceived "speaker similarity" remains one of the most common shortcomings of such systems (Yamagishi et al., 2008a).

Therefore, in the later part of this paper, we attempt to address this shortcoming, and present the results of experiments on the new RSS corpus. One possible reason that HMM-based synthetic speech sounds less like the original speaker than a concatenative system built from the same data may be the use of a vocoder, which can cause buzziness or other processing artefacts. Another reason may be that the statistical modelling itself can lead to a muffled sound, presumably due to the process of averaging many short-term spectra, which removes important detail.

In addition to these intrinsic reasons, we hypothesize that there are also extrinsic problems: some basic configuration choices in HMM synthesis have been simply taken from different fields such as speech coding, automatic speech recognition (ASR) and unit selection synthesis. For instance, 16 kHz is generally regarded as a sufficiently high waveform sampling frequency for speech recognition and synthesis because speech at this sampling frequency is intelligible to human listeners.

However speech waveforms sampled at 16 kHz still sound slightly muffled when compared to higher sampling frequencies. HMM synthesis has already demonstrated levels of intelligibility indistinguishable from natural speech (Karaiskos et al., 2008), but high-quality TTS needs also to achieve naturalness and speaker similarity.[3]

We revisited these apparently basic issues in order to discover whether current configurations are satisfactory, especially with regard to speaker similarity. As the sampling frequency increases, the differences between different auditory frequency scales such as the Mel and Bark scales (Zwicker and Scharf, 1965) implemented using a first-order all-pass function become greater. Therefore we also included a variety of different auditory scales in our experiments.

We report the results of Blizzard-style listening tests (Karaiskos et al., 2008) used to evaluate HMM-based speech synthesis using higher sampling frequencies as well as standard unit selection voices built from this corpus. The results suggest that a higher sampling frequency can have a substantial effect on HMM-based speech synthesis.

The article is organised as follows. Sections 2 and 3 give details of the RSS corpus and the Romanian front-end modules built using the Cerevoice system. In Section 4, the training procedures of the HMM-based voices using higher sampling frequencies are shown and then Section 5 presents the results of the Blizzard-style listening tests. Section 6 summarises our findings and suggests future work.

2. The Romanian speech synthesis (RSS) corpus

The Romanian speech synthesis (RSS) corpus was recorded in a hemianechoic chamber (anechoic walls and ceiling; floor partially anechoic) at the University of Edinburgh. Since the effect of microphone characteristics on HTS voices is still unknown, we used three high quality studio microphones: a Neumann u89i (large diaphragm condenser), a Sennheiser MKH 800 (small diaphragm condenser with very wide bandwidth) and a DPA 4035 (headset-mounted condenser). Fig. 1 shows the studio setup. All recordings were made at 96 kHz sampling frequency and 24 bits per sample, then downsampled to 48 kHz sampling frequency. This is a so-called over-sampling method for noise reduction. Since we oversample by a factor of 4 relative to the Nyquist rate (24 kHz) and downsample to 48 kHz, the signal-to-noise-ratio improves by a factor of 4. For recording, downsampling and bit rate conversion, we used ProTools HD hardware and software.

The speaker used for the recording is a native Romanian young female, the first author of this paper. We conducted 8 sessions over the course of a month, recording about 500 sentences in each session. At the start of each session, the speaker listened to a previously recorded sample, in order to attain a similar voice quality and intonation.

Fig. 1. Studio setup for recordings. Left microphone is a Sennheiser MKH 800 and the right one is a Neumann u89i. The headset has a DPA 4035 microphone mounted on it.

[3] Another practical, but equally important, factor is footprint. In unit selection, higher sampling frequencies may lead to a larger footprint. However, the use of higher sampling frequencies does not in itself change the footprint of a HMM-based speech synthesis system. The use of higher sampling frequencies increases computational costs for both methods.

Table 1
Phonetic coverage of each subset of the RSS corpus.

Subset	Sentences	Size [min]	Diphones	Diphones/sentence	Quinphones	Quinphones/sentence
Random	1500	104	662	0.44	41285	27.5
Diphone	1000	53	706	0.71	26385	26.3
Fairytale	1000	67	646	0.65	29484	29.4

The recording scripts comprised newspaper articles, sentences from novels, two short fairy tales written by the Romanian author Ion Creangă, and semantically unpredictable sentences (Benoît et al., 1996) intended for use in intelligibility tests. The fairy tales were divided into sentences and read in the original order of the work. Each sentence was individually presented to the speaker using a flat panel monitor.

This corpus contains disjoint training and test sets. The total recording time for the training set is about 3.5 h and it consists of about 3500 sentences: 1500 randomly chosen newspaper sentences, 1000 newspaper sentences chosen based on diphone coverage, and 1000 fairytale sentences. The recording time for the test set is about 0.5 h and it comprises 200 randomly chosen newspaper sentences, 100 randomly chosen novel sentences and 200 semantically unpredictable sentences.

Table 1 shows the total number of different diphones and quinphones in these subsets. Diphones are the typical unit used for unit selection systems and quinphones are the base unit for HMM-based speech synthesis systems.[4] A larger number of types implies that the phonetic coverage is better. From the diphones/sentence column in the table we can see that the subset designed for diphone coverage has better coverage in terms of the number of different diphone types but – looking at the quinphones/sentence column – its coverage of quinphones is slightly worse than random selection. This indicates that the appropriate text design or sentence selection policy for HMM-based speech synthesis should be different from that for unit selection.

All recorded sentences were manually endpointed and have been checked for consistency against the orthographic form. The newspaper sentences were read out using a relatively flat intonation pattern, while the fairy tales had a more narrative rhythm and prosody. Fig. 2 shows the box-plots of F_0 values extracted from all the sentences of each subset, in which the mean is represented by a solid bar across a box showing the quartiles, whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. From this figure we can see that the subset including fairy tales has wider F_0 variation than other subsets.

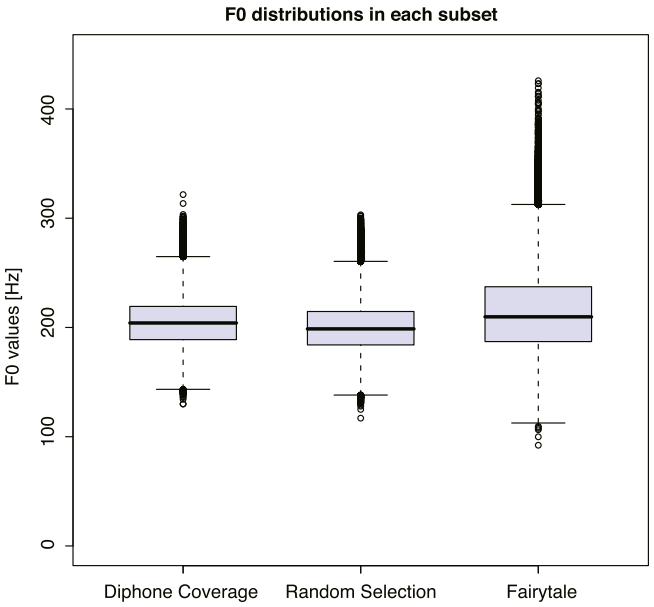


Fig. 2. F_0 distributions in each subset.

3. Romanian front-end text processing

Text processing is one of the most challenging aspects of any new language for a text-to-speech system. The great variability among different language groups and local specific alterations to standard spelling or grammar make it an important and vital part of any TTS system.

For Romanian, there are a few projects and publications regarding text processing, such as Burileanu et al. (1999), Frunza et al. (2005). However, their availability and applicability is limited. For the purpose of this study, a new text processor was developed, based on the Cerevoice development framework (Aylett and Pidcock, 2007). Language-dependent data has been gathered and probabilistic models have been trained; the front-end outputs HTS format labels comprising 53 kinds of contexts (Zen et al., 2007c). The following sections describe the resources used in developing the front-end.

3.1. Text corpus

We utilised newspaper articles obtained from the RSS feed of the Romanian free online newspaper, Adevarul. The articles were gathered over the period of August to September 2009 and they amount to about 4500 titles and over 1 million words. Due to the variety of character encodings used, the text corpus had to be cleaned and normalised before further processing.

3.2. Phonemes and letter-to-sound rules

The Romanian phonetic inventory generally consists of 7 vowels, 2 to 4 semivowels and 20 consonants. Table 2

[4] The units are further extended by adding prosodic contexts mentioned in Section 3.

Table 2
Phone set used in the experiments, given in SAMPA.

Vowel	a @ 1 e i i_0 o u
Semivowel	e_X j o w
Nasal	m n
Plosive	b d g k p t
Affricate	ts tS dZ
Fricative	f v s z S Z h
Trill	r
Approximant	l
Silence/pause	*'sil' 'pau'*

shows the phone set used in our experiments. Romanian letter-to-sound rules are straightforward. However there are several exceptions, which occur mainly in vowel sequences, such as diphthongs and triphthongs. Therefore we adopted a lightly supervised automatic learning method for letter-to-sound rules as follows: From the text corpus, the top 65,000 most frequent words were extracted. General simple initial letter-to-sound rules were written manually by a native speaker. These rules were used to phonetically transcribe the complete list of words. To deal with the exceptions above, the pronunciations of 1000 words chosen at random were checked, and corrected where necessary, by a native speaker. Using this partially-corrected dictionary of 65,000 words, letter-to-sound rules were automatically learned using a classification and regression tree (CART) (Breiman et al., 1984). The accuracy of the obtained model is about 87%, measured using 5-fold cross validation. A small additional lexicon was manually prepared to deal mainly with neologisms, whose pronunciations are typically hard to predict from spelling.

3.3. Accent

Romanian has no predefined accentual rules. Different cultural and linguistic influences cause variation in the positioning of the accent across groups of related words. However, the online SQL database of the Romanian Explicative Dictionary (DEX: http://dexonline.ro/) provides accent positioning information. Using this information from DEX directly, an accent location dictionary for the 65,000 most frequent words in the text corpus was prepared.

3.4. Syllabification

Romanian syllabification has 7 basic rules, but these can be affected by morphology, such as compound words or hyphenated compounds. These rules apply to the orthographic form of the words. In our approach, we have used the maximal onset principle applied to the phonetic transcription of the words. Onset consonant groups and vowel nuclei have been defined. Based on partial evaluation of the principle, we determined that the accuracy of the syllabification is approximately 75%. One of the major exceptions occurs in the vowel-semivowel-vowel groups, where both the vowel-semivowel and semivowel-vowel group can be a diphthong, thus a nucleus. Another important exception is represented by compound words, where the syllabification is based on morphological decomposition and not the standard rules.

3.5. Part-of-speech (POS) tagging

We used a Romanian POS tagger available online from http://www.cs.ubbcluj.ro/dtatar/nlp/WebTagger/WebTagger.htm. Most of the text corpus was split into sentences and tagged using this tool. The accuracy of the POS tagging is 70% on average, according to internal evaluation results reported by the developers of the POS tagger.

3.6. HTS labels

HTS labels were generated using the text processor, based on the recorded sentences and scripts. All the words found in the recorded sentences were checked in the lexicon for correct phonetic transcription and accent location.

4. Building HMM-based speech synthesis systems using a high sampling frequency

We adopted a recent HMM-based speech synthesis system described in (Zen et al., 2007a), which uses a set of speaker-dependent context-dependent multi-stream left-to-right state-tied (Young et al., 1994; Shinoda and Watanabe, 2000) multi-space distribution (MSD) (Tokuda et al., 2002) hidden semi-Markov models (HSMMs) (Zen et al., 2007b) that model three kinds of parameters, required to drive the STRAIGHT (Kawahara et al., 1999) mel-cepstral vocoder with mixed excitation (Kawahara et al., 2001). Once we define context-dependent labels from the language-dependent front-end outputs, the framework of this system is basically language-independent and thus we can directly use it on our data.

The sampling frequency of the speech directly affects feature extraction and the vocoder and indirectly affects HMM training via the analysis order of spectral features. The following sections give an overview of how the sampling frequency affects the first-order all-pass filter used for mel-cepstral analysis and how we can utilise higher sampling frequencies in this analysis method.

4.1. The first-order all-pass frequency-warping function

In mel-cepstral analysis (Tokuda et al., 1991), the vocal tract transfer function $H(z)$ is modelled by Mth order mel-cepstral coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\top$ as follows:

$$H(z) = \exp \boldsymbol{c}^\top \tilde{\boldsymbol{z}} = \exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m}, \quad (1)$$

where $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \ldots, \tilde{z}^{-M}]^{\top}$. $\tilde{z}^{-1}$ is defined by a first-order all-pass (bilinear) function

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{2}$$

and the warped frequency scale $\beta(\omega)$ is given as its phase response:

$$\beta(\omega) = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}. \tag{3}$$

The phase response $\beta(\omega)$ gives a good approximation to an auditory frequency scale with an appropriate choice of α.

An example of frequency warping is shown in Fig. 3. where it can be seen that, when the sampling frequency is 16 kHz, the phase response $\beta(\omega)$ provides a good approximation to the mel scale for $\alpha = 0.42$. The choice of α depends on the sampling frequency used and the auditory scale desired. The next section describes how to determine this parameter for a variety of auditory scales.

4.2. The Bark and ERB scales using the first-order all-pass function

In HMM-based speech synthesis, the mel scale is widely used. For instance, Tokuda *et al.* provide appropriate α values for the mel scale for speech sampling frequencies from 8 kHz to 22.05 kHz (Tokuda et al., 1994b).

In addition to the mel scale, the Bark and equivalent rectangular bandwidth (ERB) scales (Patterson, 1982) are also well-known auditory scales. In Smith and Abel (1999), Smith and Abel define the optimal α (in a least-squares sense) for each scale as follows:

$$\alpha_{\text{Bark}} = 0.8517\sqrt{\arctan(0.06583 f_s)} - 0.1916, \tag{4}$$

$$\alpha_{\text{ERB}} = 0.5941\sqrt{\arctan(0.1418 f_s)} + 0.03237, \tag{5}$$

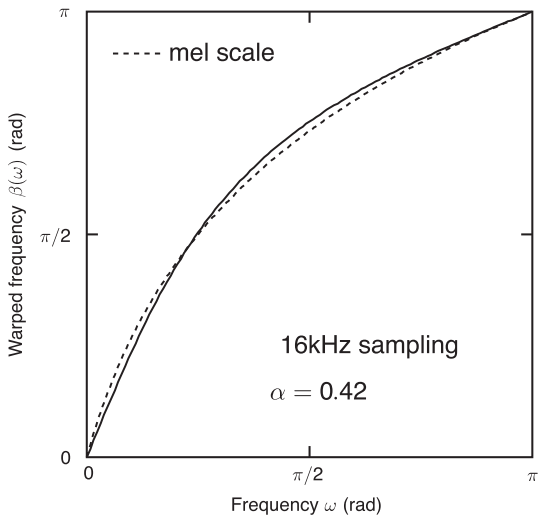


Fig. 3. Frequency warping using the all-pass function. At a sampling frequency of 16 kHz, $\alpha = 0.42$ provides a good approximation to the mel scale.

where f_s is the waveform sampling frequency. However, note that the error between the true ERB scale and all-pass scale approximated by α_{ERB} is three times larger than the error for the Bark scale using α_{Bark} (Smith and Abel, 1999). Note also that as sampling rates become higher, the accuracy of approximation using the all-pass filter becomes worse for both scales.

4.3. HMM training

The feature vector for the MSD-HSMMs consists of three kinds of parameters: the mel-cepstrum, generalised $\log F_0$ (Yamagishi and King, 2010) and a set of band-limited aperiodicity measures (Ohtani et al., 2006), plus their velocity and acceleration features.

An overview of the training stages of the HSMMs is shown in Fig. 4. First, monophone MSD-HSMMs are trained from the initial segmentation using the segmental *K*-means and EM algorithms (Dempster et al., 1977), converted to context-dependent MSD-HSMMs and re-estimated using embedded training. Then, decision-tree-based context clustering (Young et al., 1994; Shinoda and Watanabe, 2000) is applied to the HSMMs and the model parameters of the HSMMs are thus tied. The clustered HSMMs are re-estimated again using embedded training. The clustering processes are repeated until convergence of likelihood improvements (inner loop of Fig. 4) and the whole process is further repeated using segmentation labels refined with the trained models in a bootstrap fashion (outer loop of Fig. 4). In general, speech data sampled at higher rates requires a higher analysis order for mel-cepstral analysis. We therefore started by training models on lower sampling rate speech (16 kHz) with a low analysis order and gradually increased the analysis order and sampling rates via either re-segmentation of data or single-pass retraining of HMMs (Yamagishi and King, 2010).

4.4. Configurable parameters

In order to establish a benchmark system which will be useful for many future experiments, we carefully adjusted various configurable parameters as follows:

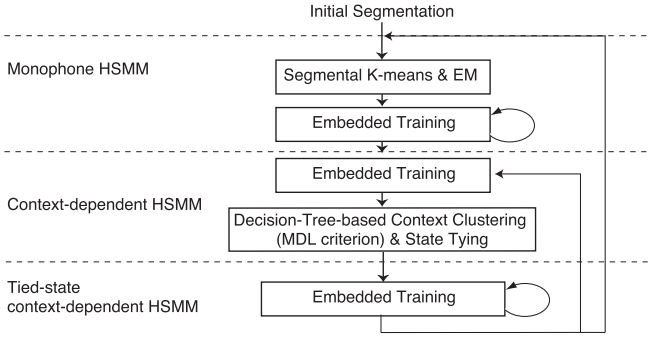


Fig. 4. Overview of HMM training stages for HTS voice building.

1. From initial analysis-by-synthesis tests using five sentences followed by informal listening, we first chose the spectral analysis method and order. Specifically, we compared mel-cepstrum and mel-generalised cepstrum (MGC) (Tokuda et al., 1994a) at orders of 50, 55, 60, 65 and 70, using Bark and ERB frequency warping scales[5] using speech data sampled at 48 kHz. The parameter to control all-pole or cepstral analysis method was set to 3 (Tokuda et al., 1994a). The results indicated the use of MGC with 60th order and the Bark scale. However, the differences between the Bark and ERB scales were found to be not as great as differences due to the sampling frequency. Our earlier research (Yamagishi and King, 2010) also found that the auditory scale – including the Mel scale – was not a significant factor. Therefore we omitted the ERB scale and the Mel scale from the listening test reported later. We repeated the same process for speech data sampled at 32 kHz and chose MGC with 44th order with the Bark scale.
2. Preliminary HMM training was then carried out to determine training data partitions. A total of 20 systems resulted from combinations of the recorded data used in sets of 500, 1000, 1500, 2500 and 3500 sentences. From informal listening, the fairy tale sentences were found to alter the overall quality of the synthesised speech, since these sentences had a more dynamic prosody than the newspaper sentences (see Fig. 2). Therefore we excluded the fairy tale set and used a 2500 sentence set in subsequent experiments.
3. We employed the data-driven generalised-logarithmic F_0 scale transform method proposed in (Yamagishi and King, 2010). The maximum likelihood estimator for the generalised logarithmic transform obtained from F0 values of all voiced frames included in the RSS database, using the optimisation method mentioned in (Yamagishi and King, 2010), was 0.333.
4. We then separated decision trees for speech from non-speech units (pauses and silences) rather than having a shared single tree.

In the experiments reported in this paper, only speech recorded using the Sennheiser MKH 800 microphone was used. Investigation of the differences caused by microphone type are left as future work.

5. Evaluation

5.1. Listening test

For the listening test, we used the framework from the Blizzard Challenge (Karaiskos et al., 2008) and evaluated speaker similarity, naturalness and intelligibility.

[5] Strictly speaking, we should call them Bark-cepstrum and ERB-cepstrum. However, for simplicity we will just call them all 'mel-cepstrum'.

We recruited a total of 54 Romanian native listeners of which 20 completed the test in purpose-built, soundproof listening booths and the rest evaluated the systems on their personal computers and audio devices, mostly using headphones. They each evaluated a total of 108 sentences randomly chosen from the test set, 36 from each category (news, novel, SUS). The speaker similarity and naturalness sections contained 18 newspaper sentences and 18 novel sentences each. 36 SUSs were used to test intelligibility.

The duration of the listening test was about 45 minutes per listener. Listeners were able to pause the evaluation at any point and continue at a later time, but the majority opted for a single listening session. Most of the listeners had rarely listened to synthetic voices; they found the judgement of naturalness and speaker similarity to be the most challenging aspects of the test.

Nine individual systems were built for the evaluation. All used the same front-end text processing. They differ in the synthesis method used (HMM-based, unit selection), sampling frequency (16 kHz, 32 kHz, 48 kHz) and the amount of data used for the training of the voice. The analysis of the three microphones is an interesting topic but, in order to make the listening tests feasible, we had to exclude this factor. The systems are identified by letter:

A Original recordings, natural speech at 48 kHz
B Unit selection system at 16 kHz, using 3500 sentences
C Unit selection system at 32 kHz, using 3500 sentences
D Unit selection system at 48 kHz, using 3500 sentences
E HMM system at 48 kHz, using 500 training sentences
F HMM system at 48 kHz, using 1500 training sentences
G HMM system at 16 kHz, using 2500 training sentences
H HMM system at 32 kHz, using 2500 training sentences
I HMM system at 48 kHz, using 2500 training sentences

By comparing systems B, C and D with E, F, G, H and I, we can see the effect of the synthesis method. By comparing systems B, C, D or G, H, I, we can see the effect of sampling frequency, per synthesis method. Comparing systems E, F, I, we can see the effect of the amount of training data for the HMMs.

In the speaker similarity task, after the listeners listened to up to 4 original recording samples, they were presented with a synthetic speech sample generated from one of the nine systems and were asked to rate similarity to the original speaker using a 5-point scale. The scale runs from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person]. In the naturalness evaluation task, listeners used a 5-point scale from 1 [Completely Unnatural] to 5 [Completely Natural]. In the intelligibility task, the listeners heard a SUS and were asked to type in what they heard. Typographical errors and spelling mistakes were allowed for in the scoring procedure. The SUS each comprised a maximum of 6 frequently-used Romanian words.

5.2. Results

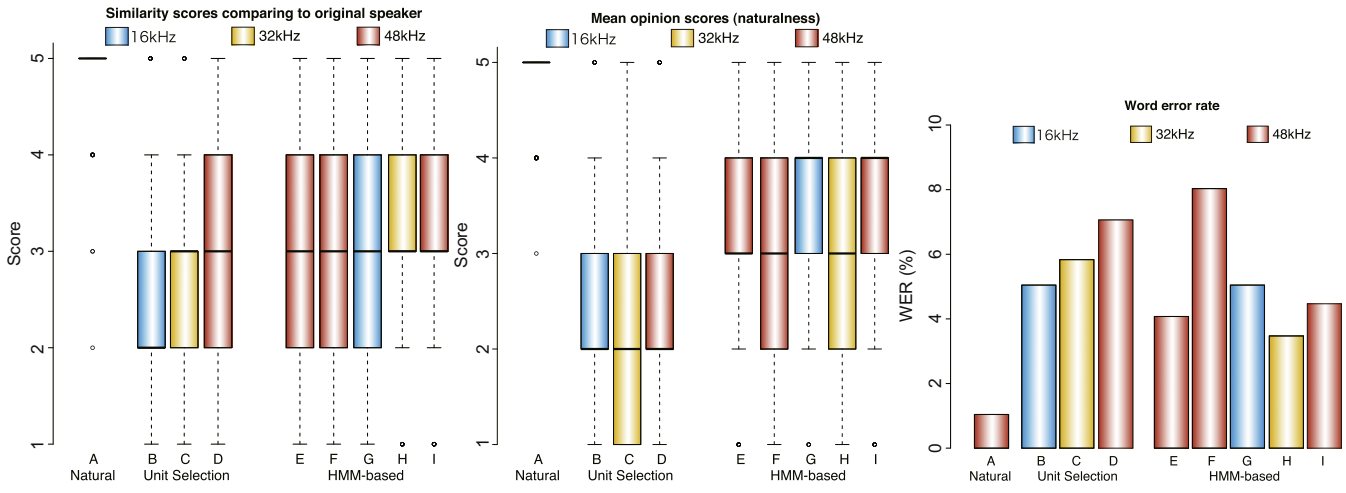
5.2.1. Speaker similarity

The left column of Fig. 5 shows the results for speaker similarity. We first observe a clear separation between the original voice (system A), HMM voices (systems E, F, G, H and I) and unit selection voices (systems B, C and D). We can also observe a clear influence of the sampling frequency over speaker similarity although improvements seem to level off at 32 kHz. This is a new and interesting finding. Also there is some influence of the amount of training data. We can see that the difference between systems E and F is less significant whereas the difference between systems F and I is significant. We believe that neither 500 nor 1500 sentences were sufficient for training models that can reproduce good speaker similarity, since the dimensionality of our features is very high due to the high order mel-cepstral analysis.

Although we expected that unit selection would have better similarity than HMM-based, the results are contrary to our expectation. This may be explained by the corpus design: In our corpus, only 1000 sentences were chosen based on diphone coverage and the remaining 2500 sentences consist of 1500 randomly chosen newspaper sentences and 1000 fairy tale sentences. Even if we combine both types of sentence, there are still 16 missing diphones and 79 diphones having fewer than 3 occurrences. Although quinphones, the base unit of HMM voices, do not have good coverage either, unit selection systems (which use diphone units) are known to be more sensitive to lack of phonetic coverage, compared to HMM-based systems (Yamagishi et al., 2008b).

5.2.2. Naturalness

We can see similar tendencies to those for the similarity task, except that sampling frequency does not seem to have any effect. The use of higher sampling frequency did not improve the naturalness of synthetic speech, in contrast to speaker similarity. This is also an interesting finding. Regarding the amount of data, we see that there are some fluctuations, although the largest amount of data typically leads to the best voice for each synthesis method.



Mean scores

A	B	C	D	E	F	G	H	I
4.9	2.6	2.5	2.7	3.1	3.0	3.1	3.4	3.3

Mean scores

A	B	C	D	E	F	G	H	I
4.8	2.2	2.5	2.4	3.3	3.0	3.4	3.4	3.4

Mean Score [**WER**%]

A	B	C	D	E	F	G	H	I
1.0	5.0	5.8	7.1	4.1	8.0	5.0	3.5	4.5

Significance at 1% level for **similarity**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	1	1	1	1	1
C	1	0	-	0	1	1	1	1	1
D	1	0	0	-	1	0	0	1	1
E	1	1	1	1	-	0	0	0	0
F	1	1	1	0	0	-	0	1	1
G	1	1	1	0	0	0	-	1	0
H	1	1	1	1	0	1	1	-	0
I	1	1	1	1	0	1	0	0	-

Significance at 1% level for **naturalness**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	1	1	1	1	1
B	1	-	0	0	1	1	1	1	1
C	1	0	-	0	1	1	1	1	1
D	1	0	0	-	1	1	1	1	1
E	1	1	1	1	-	1	0	0	0
F	1	1	1	1	1	-	1	1	1
G	1	1	1	1	0	1	-	0	0
H	1	1	1	1	0	1	0	-	0
I	1	1	1	1	0	1	0	0	-

Significance at 1% level for **WER**

	A	B	C	D	E	F	G	H	I
A	-	1	1	1	0	1	1	0	1
B	1	-	0	0	0	0	0	0	0
C	1	0	-	0	0	0	0	0	0
D	1	0	0	-	0	0	0	0	0
E	0	0	0	0	-	0	0	0	0
F	1	0	0	0	0	-	0	0	0
G	1	0	0	0	0	0	-	0	0
H	0	0	0	0	0	0	0	-	0
I	1	0	0	0	0	0	0	0	-

Fig. 5. Listening tests results. There are three columns of plots and tables which are, from left to right, similarity to original speaker, mean opinion score for naturalness, and intelligibility. The similarity and naturalness plots on the upper row are box plots where the median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range. The three tables in the middle row give the mean scores of each system. The tables in the bottom row indicate significant differences between pairs of systems, based on Wilcoxon signed rank tests with alpha Bonferroni corrected (1% level); '1' indicates a significant difference.

5.2.3. Intelligibility

Unfortunately there appears to be something of a ceiling effect on intelligibility. Absolute values of WER are generally small: both synthesis methods have good intelligibility. Even though we observe that systems D and F have a slightly higher error rate, there are no statistically significant differences between any pairs of synthetic voices in terms of WER. To confirm this we performed a small additional test including paronyms and obtained the same results. We believe that the lack of significant differences between systems is partly caused by the nature of the simple grapheme-to-phoneme rules in Romanian. Even for SUSs and paronyms, both natural and synthetic speech are easy to transcribe, leading to WERs close to zero. This result suggests there is a need for better evaluation methods for the intelligibility of synthetic speech in languages such as Romanian.

5.2.4. Listening environments

We performed an ANOVA test to discover whether the listening environment affects the results. An ANOVA test at 1% significance level shows that only the system C (unit selection system at 32 kHz, using 3500 sentences) in the similarity test was affected by the listening environment. The subjects who completed the test in the listening booths generally gave lower similarity scores for system C.

5.2.5. Summary

This RSS corpus is probably better suited to HMM-based synthesis than to unit selection. All speech synthesis systems built using the corpus have good intelligibility. However, we need to design a better evaluation of the system's intelligibility in simple grapheme-to-phoneme languages such as Romanian.

We found that the sampling frequency is an important factor for speaker similarity. More specifically, downsampling speech data in this corpus to 32 kHz does no harm, but downsampling to 16 kHz degrades speaker similarity. The use of higher sampling frequency, however, did not improve either the naturalness or intelligibility of synthetic speech.

These results are consistent with existing findings: (Fant, 2005) mentions that almost all the linguistic information from speech is in the frequency range 0 to 8 kHz. This implies that a 16 kHz sampling frequency (and thus 8 kHz Nyquist frequency) is sufficient to convey the linguistic information. Our results also shown that using sampling frequencies over 16 kHz did not improve the intelligibility of synthetic speech. On the other hand, a classic paper regarding sampling frequency standardisation (Muraoka et al., 1978) reported that a cut-off frequency of less than 15 kHz may deteriorate audio quality. This means that the sampling frequency used should be higher than 30 kHz. In fact, our results do show that downsampling to 16 kHz degrades speaker similarity. Therefore we can conclude that the naturalness and intelligibility of synthetic speech only require transmission of linguistic information, which can be achieved at 16 kHz sampling frequency, whereas speaker similarity of synthetic speech is affected by audio quality (requiring a higher sampling rate).

5.3. Demos

We encourage interested readers to listen to audio samples comprising some of the materials used for listening tests http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/rss.html and the first 3 chapters of a Romanian public-domain novel "Moara cu noroc" by Ioan Slavici, available online via http://octopus.utcluj.ro:56337/moaraCuNoroc/moaraCuNoroc.rss. We also encourage them to test our live demo http://octopus.utcluj.ro:56337/HTS_Romanian-Demo/index.php. The RSS database itself can be downloaded from http://octopus.utcluj.ro:56337/RORelease/.

6. Conclusions

This paper has introduced a newly-recorded high-quality Romanian speech database which we call "RSS", along with Romanian front-end modules and HMM-based voices. In order to promote Romanian speech technology research, all of these resources are freely available for academic use.

From the listening tests presented here, we conclude that (1) the RSS corpus is well-suited for HMM-based speech synthesis and (2) that the speech synthesis systems built from the corpus have good intelligibility.

Using the RSS corpus, we have also revisited some basic configuration choices made in HMM-based speech synthesis such as the sampling frequency and auditory scale, which have been typically chosen based on experience from other fields. We found that higher sampling frequencies (above 16 kHz) improved speaker similarity. More specifically, the speech data in this corpus can be downsampled to 32 kHz without affecting results but that downsampling to 16 kHz degrades speaker similarity.

Future work includes an analysis of each of the three microphones used and designing a better intelligibility evaluation for the simple grapheme-to-phoneme languages, such as Romanian.

Acknowledgements

A simplified description of some of this research was published in (Yamagishi and King, 2010).

Adriana Stan is funded by the European Social Fund, project POSDRU/6/1.5/S/5 and was visiting CSTR at the time of this work. Junichi Yamagishi and Simon King are partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant agreement 213845 (the EMIME project).

This work has 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).

We would like to thank CereProc staff for support with the text processing tools, Catalin Francu for assistance with the DEX-online database, and the authors of the Romanian POS Tagger. The first author would also like to thank everyone at CSTR – especially Oliver Watts – for their support and guidance.

References

Aylett, M., Pidcock, C., 2007. The CereVoice characterful speech synthesiser SDK. Proc. AISB 2007. Newcastle, U.K., pp. 174–178.

Benoît, C., Grice, M., Hazan, V., 1996. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Comm. 18 (4), 381–392.

Black, A., Cambpbell, N., 1995. Optimising selection of units from speech database for concatenative synthesis. In: Proceedings of EUROSPEECH-95, pp. 581–584.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Burileanu, D., Dan, C., Sima, M., Burileanu, C., 1999. A parser-based text preprocessor for Romanian language TTS synthesis. In: Proc. EUROSPEECH-99. Budapest, Hungary, pp. 2063–2066.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc. Ser. B 39 (1), 1–38.

Fant, G., 2005. Speech acoustics and phonetics: selected writings. Chapter Speech Perception. Springer, Netherlands, pp. 199–220.

Ferencz, A., 1997. Contributii la dezvoltarea sintezei text-vorbire pentru limba romana. Ph.D. thesis, University of Cluj-Napoca.

Frunza, O., Inkpen, D., Nadeau, D., 2005. A text processing tool for the Romanian language. In: Proceedings of EuroLAN 2005: Workshop on Cross-Language Knowledge Induction, Cluj-Napoca.

Hunt, A. and Black, A., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP-96, pp. 373–376.

Karaiskos, V., King, S., Clark, R.A.J., Mayo, C., 2008. The Blizzard Challenge 2008. In: Proceedings of Blizzard Challenge Workshop, Brisbane, Australia.

Kawahara, H., Masuda-Katsuse, I., Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Comm. 27, 187–207.

Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: 2nd MAVEBA.

Muraoka, T., Yamada, Y., Yamazaki, M., 1978. Sampling-frequency considerations in digital audio. J. Audio Eng. Soc. 26 (4), 252–256.

Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2006. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In: Proceedings of Interspeech 2006, pp. 2266–2269.

Patterson, R., 1982. Auditory filter shapes derived with noise stimuli. J. Acous. Soc. Amer. 76, 640–654.

Shinoda, K., Watanabe, T., 2000. MDL-based context-dependent subword modeling for speech recognition. J. Acous. Soc. Jpn. (E) 21, 79–86.

Smith III, J.O., Abel, J.S., 1999. Bark and ERB bilinear transforms. IEEE Trans. Speech Audio Process. 7 (6), 697–708.

Tokuda, K., Kobayashi, T., Fukada, T., Saito, H., Imai, S., 1991. Spectral estimation of speech based on mel-cepstral representation. IE ICE Trans. Fundam. J74-A (8), 1240–1248 (in Japanese).

Tokuda, K., Kobayashi, T., Masuko, T., Imai, S., 1994a. Mel-generalized cepstral analysis – a unified approach to speech spectral estimation. In: Proceedings of ICSLP-94, Yokohama, Japan, pp. 1043–1046.

Tokuda, K., Kobayashi, T., Imai, S., 1994b. Recursive calculation of mel-cepstrum from LP coefficients. In: Technical Report of Nagoya Institute of Technology.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 2002. Multi-space probability distribution HMM. IEICE Trans. Inf. & Syst. E85-D (3), 455–464.

Yamagishi, J., King, S., 2010. Simple methods for improving speaker-similarity of HMM-based speech synthesis. In: Proceedings of ICASSP 2010. Dallas, TX, pp. 4610–4613.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., Tokuda, K., 2008a. The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In: Proceedings of Blizzard Challenge 2008, Brisbane, Australia.

Yamagishi, J., Ling, Z., King, S., 2008b. Robustness of HMM-based speech synthesis. In: Proceedings of Interspeech 2008. Brisbane, Australia, pp. 581–584.

Young, S., Odell, J., Woodland, P., 1994. Tree-based state tying for high accuracy acoustic modeling. In: Proceedings of ARPA Human Language Technology Workshop, pp. 307–312.

Zen, H., Toda, T., Nakamura, M., Tokuda, K., 2007a. Details of Nitech IEICE Trans, HMM-based speech synthesis system for the Blizzard Challenge 2005. Inf. & Syst. E90-D (1), 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2007b. A hidden semi-Markov model-based speech synthesis system. IEICE Trans. Inf. & Syst. E90-D (5), 825–834.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Tokuda, K., 2007c. The HMM-based speech synthesis system (HTS) version 2.0. In: Proceedings of Sixth ISCA Workshop on Speech Synthesis, pp. 294–299.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Comm. 51 (11), 1039–1064.

Zwicker, E., Scharf, B., 1965. A model of loudness summation. Psych. Rev. 72, 2–26.

Interactive Intonation Optimisation Using CMA-ES and DCT Parameterisation of the F0 Contour for Speech Synthesis

Adriana STAN, Florin-Claudiu POP, Marcel CREMENE,
Mircea GIURGIU, Denis PALLEZ

Abstract [1] Expressive speech is one of the latest concerns of text-to-speech systems. Due to the subjectivity of expression and emotion realisation in speech, humans cannot objectively determine if one system is more expressive than the other. Most of the text-to-speech systems have a rather flat intonation and do not provide the option of changing the output speech. We therefore present an interactive intonation optimisation method based on the pitch contour parameterisation and evolution strategies. The Discrete Cosine Transform (DCT) is applied to the phrase level pitch contour. Then, the genome is encoded as a vector that contains 7 most significant DCT coefficients. Based on this initial individual, new speech samples are obtained using an interactive Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm. We evaluate a series of parameters involved in the process, such as the initial standard deviation, population size, the dynamic expansion of the pitch over

Adriana STAN
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Adriana.Stan@com.utcluj.ro

Florin-Claudiu POP
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Florin.Pop@com.utcluj.ro

Marcel CREMENE
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Cremene@com.utcluj.ro

Mircea GIURGIU
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Mircea.Giurgiu@com.utcluj.ro

Denis PALLEZ
Laboratoire d'Informatique, Signaux, et Systèmes de Sophia-Antipolis (I3S), Université de Nice Sophia-Antipolis, France, e-mail: Denis.Pallez@unice.fr

[1] **Accepted for publication in In Proceedings of the 5_{th} Workshop on Nature Inspired Cooperative Strategies for Optimisation, in series Studies in Computational Intelligence, vol. 387, Springer, 2011, http://www.springer.com/engineering/computational+intelligence+and+complexity/book/978-3-642-24093-5**

the generations and the naturalness and expressivity of the resulted individuals. The results have been evaluated on a Romanian parametric-based speech synthesiser and provide the guidelines for the setup of an interactive optimisation system, in which the users can subjectively select the individual which best suits their expectations with minimum amount of fatigue.

1 Introduction

Over the last decade text-to-speech (TTS) systems have evolved to a point where in certain scenarios, non-expert listeners cannot distinguish between human and synthetic voices with 100% accuracy. One problem still arises when trying to obtain a natural, more expressive sounding voice. Several methods have been applied ([17], [20]), some of which have had more success than others and all of which include intonation modelling as one of the key aspects. Intonation modelling refers to the manipulation of the pitch or fundamental frequency (F0). The expressivity of speech is usually attributed to a dynamic range of pitch values. But in the design of any speech synthesis system (both concatenative and parameteric), one important requirement is the flat intonation of the speech corpus, leaving limited options for the synthesised pitch contours.

In this paper we propose an interactive intonation optimisation method based on evolution strategies. Given the output of a synthesiser, the user can opt for a further enhancement of its intonation. The system then evaluates the initial pitch contour and outputs a small number of different versions of the same utterance. Provided the user subjectively selects the best individual in each set, the next generation is built starting from this selection. The *dialogue* stops when the user considers one of a generation's individual satisfactory. The solution for the pitch parameterisation is the Discrete Cosine Transform (DCT) and for the interactive step, the Covariance Matrix Adaptation-Evolution Strategy (CMA-ES).

This method is useful in the situation where non-expert users would like to change the output of a speech synthesiser to their preference. Also, under resourced languages or limited availability of speech corpora could benefit from such a method. The prosodic enhancements selected by the user could provide long-term feedback for the developer or could lead to a *user-adaptive* speech synthesis system.

1.1 Problem statement

In this subsection we emphasise some aspects of the current state-of-the-art speech synthesisers which limit the expressiveness of the result:

Issue #1: Some of the best TTS systems benefit from the prior acquisition of a large speech corpus and in some cases extensive hand labelling and rule-based

intonation. But this implies a large amount of effort and resources, which are not available for the majority of languages.

Issue #2: Most of the current TTS systems provide the user with a single unchangeable result which can sometimes lack the emphasis or expressivity the user might have hoped for.

Issue #3: If the results of a system can be improved, it usually implies either additional annotation of the text or a trained specialist required to rebuild most or all of the synthesis system.

Issue #4: Lately, there have been studies concerning more objective evaluations of the speech synthesis, but in the end the human is the one to evaluate the result and this is done in a purely subjective manner.

1.2 Related work

To the best of our knowledge, evolution strategies have not been previously applied to speech synthesis. However, the related genetic algorithms have been used in articulatory [1] or neural networks based [11] speech synthesisers. A study of interactive genetic algorithms applied to emotional speech synthesis is presented in [8]. The authors use the XML annotation of prosody in Microsoft Speech SDK and try to convert neutral speech to one of the six basic emotions: *happiness, anger, fear, disgust, surprise* and *sadness*. The XML tags of the synthesised speech comprise the genome. Listeners are asked to select among 10 speech samples at each generation and to stop when they consider the emotion in one of the speech samples consistent with the desired one. The results are then compared with an expert emotional speech synthesis system. Interactive evolutionary computation has, on the other hand, been applied to music synthesis [10],and music composition [3], [9].

2 DCT parameterisation of the F0 Contour

In text-to-speech one of the greatest challenges remains the intonation modelling. There are many methods proposed in order to solve this problem, some taking into account a phonological model [2], [15] and others simply parameterising the pitch as a curve [18]. Curve parameterisation is a more efficient method in the sense that no manual annotation of the text to be synthesised is needed and thus not prone to subjectivity errors.

Because in this study we are not using prior text annotations or additional information, we chose a parameterisation based on the DCT, that partially adresses Issue #1 of the Problem Statement section.

DCT is a discrete transform which expresses a sequence of discrete points as a sum of cosine functions oscillating at different frequencies with zero phase. The

are several forms, but the most common one is DCT-II (Eq. 1). The coefficients are computed according to Eq. 2.

$$X_k = \frac{1}{2}x_0 + \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right], k = 0, 1...N-1 \quad (1)$$

$$c_n = \sum_{x=0}^{M-1} s(x) cos\left[\frac{x}{M}n(x+\frac{1}{2})\right] \quad (2)$$

DCT applied to pitch parameterisation has been extensively studied in [7], [13] and [19]. These works prove that DCT is an efficient way to parameterise the pitch with minimum error. Also, the principle behind DCT adheres to the superpositional aspect [14] of the fundamental frequency. The principle states that the pitch can be broken down into separate layers of realisation, heuristically named phrase, word, syllable and phoneme, in the sense that the cognitive process of speech derives a phrase contour unto which the rest of the layers are overlapped. Another important aspect of the DCT is its direct inverse transform. This is needed in the re-synthesis of the pitch contour from the DCT coefficients (Eq. 1).

The method we propose addresses the issue of modelling the phrase level intonation, or trend. Starting from a flat intonation, we would like to derive more dynamic and expressive contours. Therefore, we consider the phrase layer to be represented by the inverse DCT transform of the DCT1 to DCT7 coefficients of the pitch DCT. This assumption is also supported by the results presented in [19]. DCT0 represents the mean of the curve and in our case it is speaker dependent. Using DCT0 in the genome encoding would undesirably change the pitch of the speaker, our focus being on the overall trend of the phrase intonation. The phrase level is then subtracted from the overall contour, and the result is retained and will be referred to as *high level pitch information*. Fig. 1 presents an example of a pitch contour, the phrase level contour based on the inverse DCT of the DCT1-DCT7 coefficients and the high level pitch information. It can be observed that the phrase level contour represents the relative trend of the voiced segments intonation, while the high level information has a relatively flat contour with variations given by the word, syllable and phoneme levels.

Because DCT cannot parameterise fast variations with a small number of coefficients, the unvoiced segments of the F0 contour were interpolated using a cubic function (Eq. 3). During the interactive step we apply the inverse DCT transform over the winner's genome, add the high level pitch information and synthesise the speech using the resulted F0 contour.

$$f(x) = ax^3 + bx^2 + cx + d \quad (3)$$

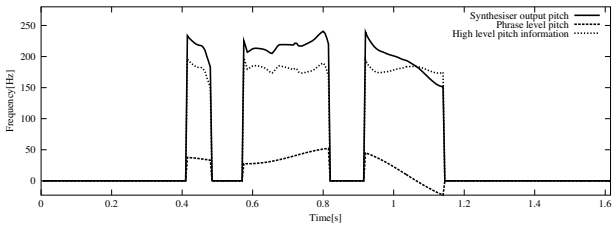


Fig. 1 An example of a pitch contour for the utterance "Ce mai faci" ("How are you"), the phrase level contour based on the inverse DCT of DCT1-DCT7 coefficients and the high level pitch information.

3 Optimisation using CMA-ES

CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) was proposed by Hansen and Ostermeier [5] as an evolutionary algorithm to solve unconstrained or bounded constraint, non-linear optimisation problems defined in a continuous domain. In an evolutionary algorithm, a *population* of genetic representations of the solution space, called *individuals*, is updated over a series of iterations, called *generations*. At each generation, the best individuals are selected as *parents* for the next generation. The function used to evaluate individuals is called the *fitness* function.

The search space is explored according to the genetic operations used to update the individuals in the parent population and generate new offspring. In the case of evolution strategy (ES), the selection and mutation operators are primarily used, in contrast to the genetic algorithm (GA) proposed by Holland [6], which considers a third operator – crossover. Also, in GA the number of mutated genes per individual is determined by the *mutation probability*, while in ES mutation is applied to all genes, slightly and at random.

If mutation is according to a multivariate normal distribution of mean m and covariance matrix C, then CMA-ES is a method to estimate C in order to minimise the search cost (number of evaluations). First, for the mean vector $m \in \mathbb{R}^n$, which is assimilated to the preferred solution, new individuals are sampled according to the normal distribution described by $C \in \mathbb{R}^{n \times n}$:

$$x_i = m + \sigma y_i \tag{4}$$
$$y_i \sim N_i(0, C), i = 1..\lambda$$

where λ is the size of the offspring population and $\sigma \in \mathbb{R}_+$ is the step size.

Second, sampled individuals are evaluated using the defined fitness function and the new population is selected. There are two widely used strategies for selection: $(\mu + \lambda)$-ES and (μ, λ)-ES, where μ represents the size of the parent population. In $(\mu + \lambda)$-ES, to keep the population constant, the λ worst individuals are discarded

after the sampling process. In (μ, λ)-ES all the parent individuals are discarded from the new population in favour of the λ new offspring.

Third, m, C and σ are updated. In the case of (μ, λ)-ES, which is the strategy we chose to implement our solution, the new mean is calculated as follows:

$$m = \sum_{i=1}^{\mu} w_i x_i \tag{5}$$

$$w_1 \geq .. \geq w_\mu, \sum_{i=1}^{\mu} w_i = 1$$

where x_i is the i-th ranked solution vector ($f(x_1) \leq .. \leq f(x_\lambda)$) and w_i is the weight for sample x_i.

The covariance matrix C determines the shape of the distribution ellipsoid and it is updated to increase the likelihood of previously successful steps. Details about updating C and σ can be found in [4] .

CMA-ES is the proposed solution for Issues #2, #3 and #4 through the generation of several individuals (i.e. speech samples) the user can chose from, the extension of the coefficients' space and the subjective fitness function for the interactive step.

4 Proposed solution

Combining the potential of the DCT parameterisation and evolution strategies, we introduce an interactive solution for the intonation optimisation problem, which requires no previous specific knowledge of speech technology. To achieve this, three problems need to be solved: *1)* generate relevant synthetic speech samples for a user to chose from, *2)* minimise user fatigue and *3)* apply the user feedback to improve the intonation of the utterance.

We solve the first problem by using CMA-ES to generate different speech samples, normally distributed around the baseline output of a Romanian speech synthesis system [16] based on HTS (Hidden Markov Models Speech Synthesis System) [21]. We consider a *genome* encoded using a vector of 7 genes, where each gene stores the value of a DCT coefficient, from DCT1 to DCT7. We start with an initial mean vector m that stores the DCT coefficients of the F0 phrase level generated by the HTS system and an initial covariance matrix $C = I \in \mathbb{R}^{7\times7}$. In each generation, new individuals are sampled according to Eq. (4).

In the next step, the user needs to evaluate generated individuals. If the population size is too large, the user may get tired before a suitable individual is found or might not spot significant differences between the individuals. On the other hand, if the population size is too small and the search space is not properly explored, a suitable individual may not be found. CMA-ES is known to converge faster even with smaller population than other evolutionary algorithms, but it was not previously applied to solve interactive problems. On the other hand, interactive genetic algo-

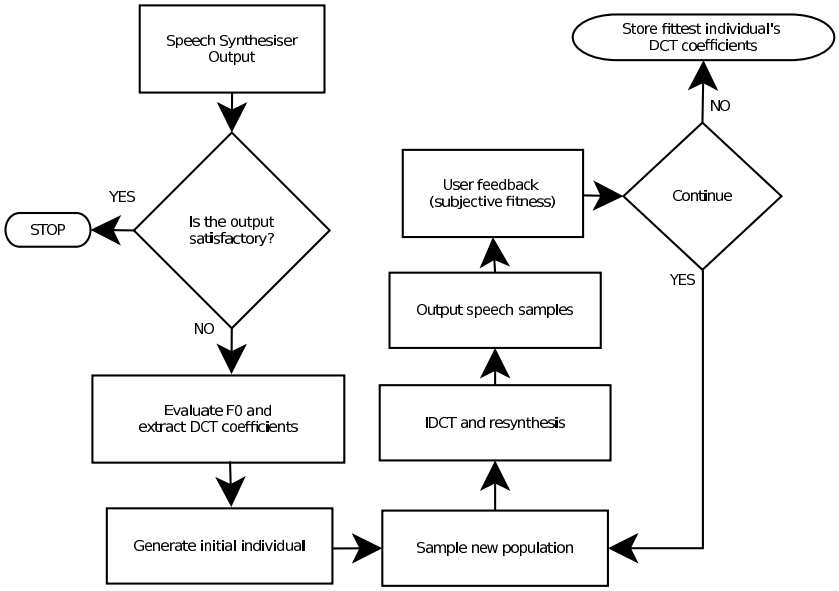


Fig. 2 Proposed method flow chart

rithms (IGA) have been extensively studied, but do not converge as fast as CMA-ES for non-linear non-convex problems. Faster convergence means fewer evaluations, therefore reducing user fatigue.

For the interactive version of CMA-ES, we used a *single elimination tournament* fitness [12]. In this case, the individuals are paired at random and play one game per pair. Losers of the game are eliminated from the tournament. The process repeats until a single champion is left. The fitness value of each individual is equal to the number of played games. Each pair of individuals is presented to the user in the form of two speech samples. Being a subjective evaluation, the choice would best suit the user's requirements, thus giving the winner of a population.

The fitness value is used by CMA-ES to update mean vector m, the covariance matrix C and the standard deviation σ. A new population of individuals is sampled based on the updated values and the process repeats. The flow chart of the proposed method is presented in Fig. 2.

5 Results

The results presented below focus on establishing the correct scenario for the interactive application and on the ease of use on behalf of the listeners/users. This

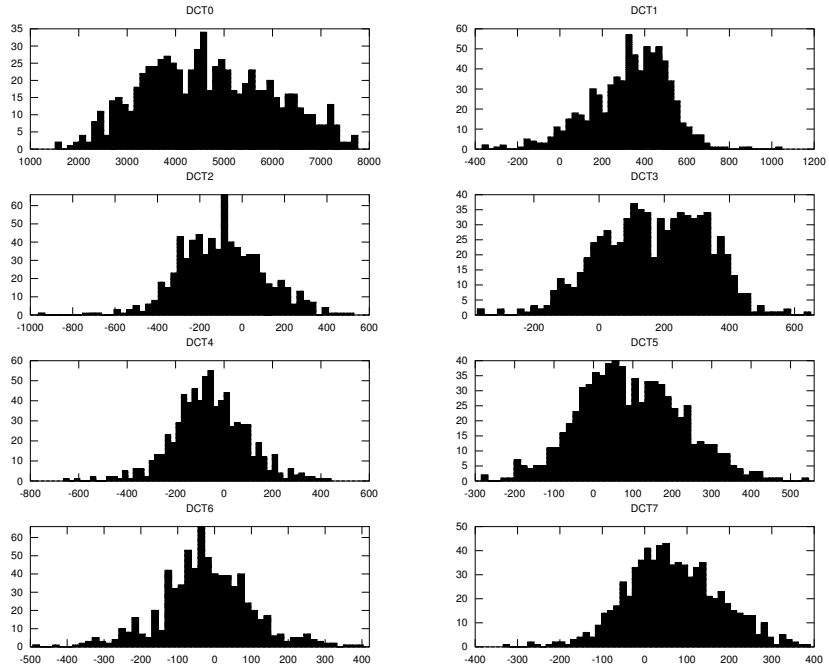


Fig. 3 The histograms of the first 8 DCT coefficients of the rnd1 subset of the RSS speech corpus. The 0x axis represents the values of the DCT coefficients separated in 50 equally spaced bins. The 0y axis is the number of coefficients equal to the values within the domain bin.

implies the evaluation of several parameters involved, such as: *initial standard deviation of the population* – gives the amount of dynamic expansion of pitch –, *the population size* – determines the number of samples the user has to evaluate in each generation, *the expressivity and naturalness of the generated individuals* – assures correct values for the pitch contour.

As a preliminary step in defining the standard deviation of the population, we employed an analysis of all the DCT coefficients within the *rnd1* subset of the Romanian Speech Synthesis corpus [16]. *rnd1* comprises 500 newspaper sentences read by a native Romanian female speaker. The number of phrases within this subset is 730 with an average length of 1.7 seconds. The intonation of the speech is flat, declarative. The histograms of the first 8 DCT coefficients of the phrases in *rnd1* are presented in Fig. 3. We included DCT0 as well for an overall view as it represents the mean of the pitch contour and it is speaker dependent. This coefficient was not used in the estimation of the phrase level contour. The means and standard deviations of the coefficients are presented in Table 1. The average pitch contour resulted from the mean values of the DCT coefficients and the average duration of the *rnd1* subset is shown in Fig. 4.

Table 1 Means and standard deviation of the DCT coefficients in *rnd1* subset with corresponding variations in Hz for an average length of 1.7 seconds.

'Coefficient	Mean	Mean F0 [Hz]	Standard deviation	Maximum F0 deviation [Hz]	
				- 1 std dev	+1 std dev
DCT0	4690.300	251-257	1318.300	179-186	322-329
DCT1	331.750	± 4	185.850	±12	±40
DCT2	-95.087	±7	197.470	±22	±7
DCT3	168.270	±12	161.030	±0.55	±25
DCT4	-57.100	±4	151.600	±16	±7
DCT5	94.427	±7	130.150	±2	±17
DCT6	-22.312	±1	123.020	±11	±7
DCT7	67.095	±5	110.370	±3	±13

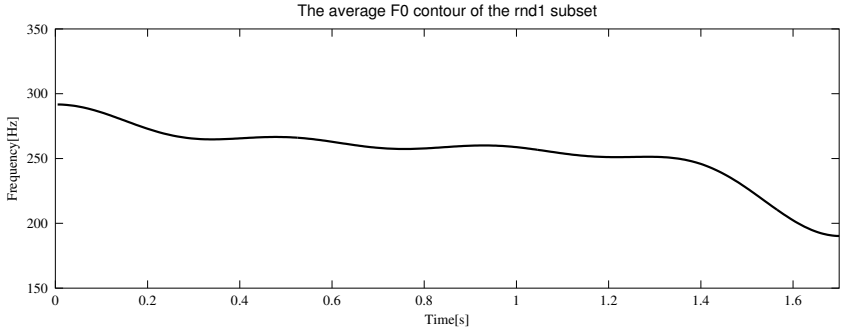


Fig. 4 The average pitch contour resulted from the mean values of the DCT0-DCT7 coefficients for the average length of 1.7 seconds in the *rnd1* subset.

DCT1 has the most important influence in the F0 contour after DCT0. The mean value of the DCT1 coefficient is 331.75 with a standard deviation of 185.85 and the maximum F0 variation is given by the *+1 std. dev.* (i.e. 331.75+185.85 = 517.6) of around 40 Hz. One of the issues addressed in this paper is the expansion of the pitch range. This means that having a standard deviation of the flat intonation speech corpus, we should impose a higher value for it while generating new speech samples, but it should not go up to the point where the generated pitch contours contain F0 values which are not natural. In Fig. 5 we compare the third generation for an initial standard deviation of 150 and 350 respectively. We can observe in the 350 case that individual 3 has F0 values going as low as 50 Hz – unnatural, while for a standard deviation of 150, the F0 contours do not vary too much from the original one and lead to a less dynamic output. Given these results, we selected a standard deviation of 250. An important aspect to be noticed from Table 1 is that all the 7 coefficients have approximately the same standard deviation. This means that

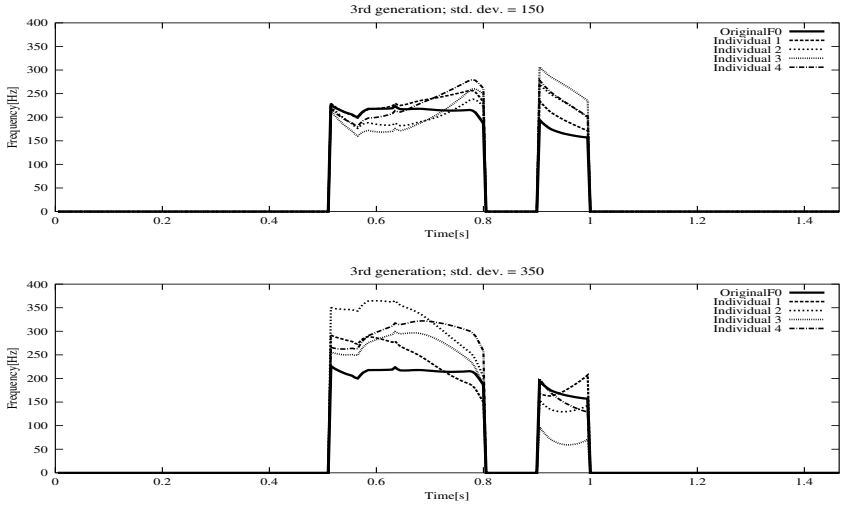


Fig. 5 The 3rd generation population of the F0 contour for the phrase "Ce mai faci?" ("How are you?"), with an initial standard deviation of 150 and 350 respectively. Original F0 represents the pitch contour produced by the synthesiser.

imposing a variation based on DCT1 does not exceed natural values for the rest of the coefficients.

The single elimination tournament fitness we used to evaluate the individuals requires the user to provide feedback for $n-1$ games, where n is the population size. So that the population size has a great importance in setting up the interactive application. Several values have been selected for it and the results are shown in Fig. 6. Although the highest the number of individuals the more samples the user can choose from, this is not necessarily a good thing in the context of user fatigue. But having only 2 individuals does not offer enough options for the user to choose from. We therefore suggest the use of 4 individuals per generation as a compromise between sample variability and user fatigue.

Another evaluation is the observation of the modification of the pitch contour from one generation to the other. Fig. 7 presents the variation of F0 from the initial population to the third. It can be observed that starting with a rather flat contour, by the third generation the dynamics of the pitch are much more expanded, resulting a higher intonation variability within and between generations. It is also interesting to observe the phrase level contours (Fig. 8). This is a more relevant evaluation as it shows the different trends generated by CMA-ES and the trend selected by the user in each generation. The selected trend can be used in the adaptation of the overall synthesis. In our example, the user selected an intonation with a high starting point and a descending slope afterwards, while another user could have chosen individual 1 which contains an initial ascending slope.

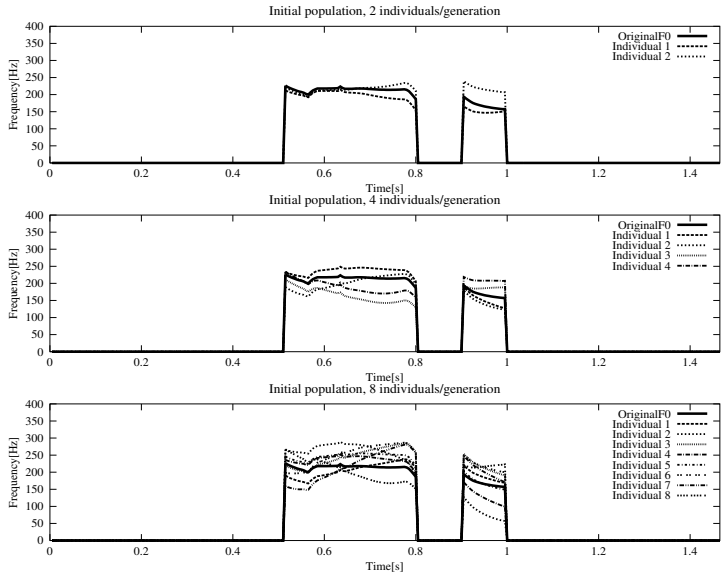


Fig. 6 Variation in the population size. Phrase "Ce mai faci?" ("How are you?"). Original F0 represents the pitch contour produced by the synthesiser.

In order to establish the naturalness of the generated individuals and the enhanced expressivity of the winners of each generation, a small listening test was conducted. At first, a user was asked to select the winners over 4 generations for 10 phrases. Initial standard deviation was 250 and with a population size of 4. Then 10 listeners had to attribute Mean Opinion Scores (MOS) for the samples in two categories: *Naturalness* – the generated samples were compared to original recordings on a scale of [1 - Unnatural] to [5 - Natural]. All the individuals of the four generations were presented. *Expressivity* – the winners of each generation were compared to the correspondent synthesised versions of them. The listeners had to mark on a scale of [1-Less expressive] to [5-More expressive] the generated samples in comparison to the synthesiser's output. The results of the test are presented in Fig. 9. In the naturalness test, all the generations achieved a relatively high MOS score, with some minor differences for the 4^{th} generation. The expressivity test reveals the fact that all the winning samples are more expressive than the originally synthesised one. The test preliminary conclude the advantages of this method. While maintaining the naturalness of the speech, its expressivity is enhanced.

Examples of speech samples generated by our method can be found at `http://www.romaniantts.com/nicso2011`.

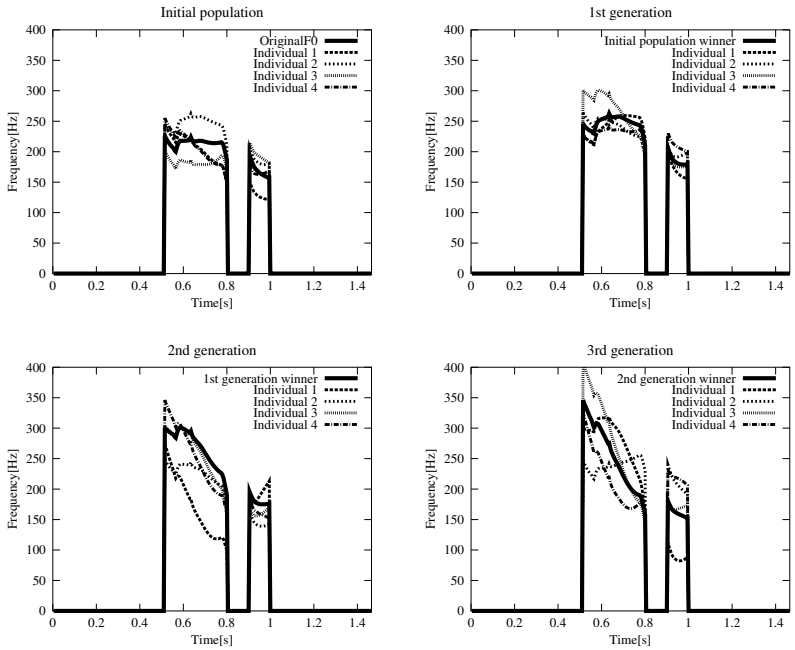


Fig. 7 Evolution of the F0 contour over 3 generations, standard deviation = 250, phrase "Ce mai faci?" ("How are you?"). Original F0 represents the pitch contour produced by the synthesiser.

6 Conclusions

We introduced a new method for intonation optimisation of a speech synthesis system based on CMA-ES and DCT parameterisation of the pitch contour. The interactive manner of the optimisation allows the users to select an output which best suits their expectations. The novelty of the solution consists in using no prosodic annotations of the text, no deterministic rules and no predefined speaking styles. Also, to the best of our knowledge, this is one of the first applications of CMA-ES for an interactive problem.

The evaluation of the system's parameters provide the guidelines of the setup for an interactive application. The proposed solutions ensure an optimal value for standard deviation and population size in order to concurrently maintain the naturalness of the speech samples, while expanding the dynamics of the pitch. The latter indicators have been evaluated in the listening test. The listening test also determined the enhancement of the expressivity of the samples.

One drawback to our 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

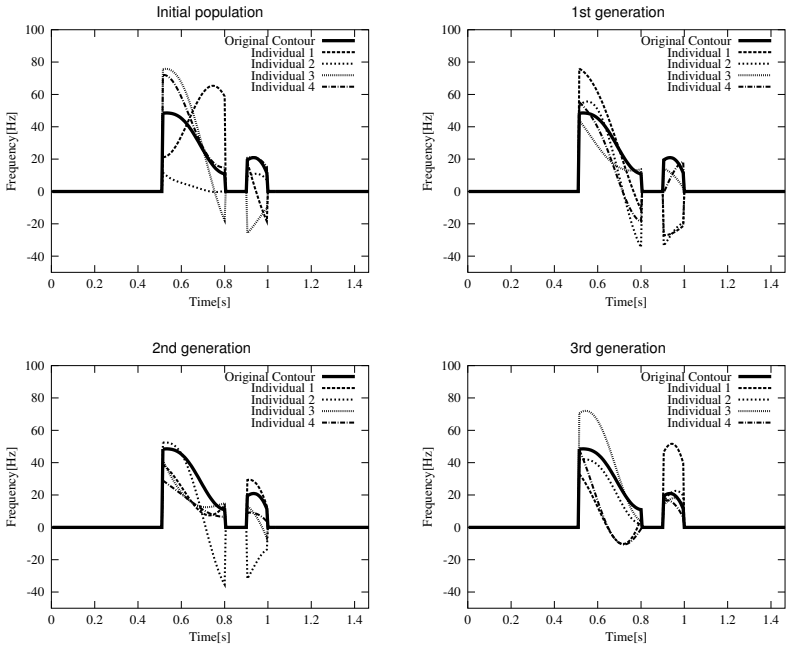


Fig. 8 Evolution of the phrase contour trend over 3 generations for the utterance "Ce mai faci" ("How are you"). Original contour represents the pitch contour produced by the synthesiser.

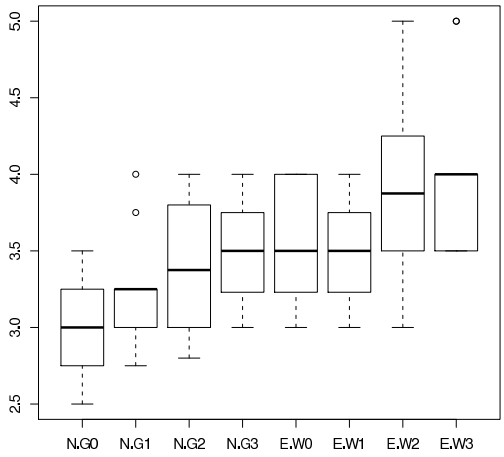


Fig. 9 Box plots results of the listening test. N-G*x* represent the results for the naturalness test of each generation and E-G*x* represent the results for the expressivity test of each generation. The median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range.

standard deviation is similar and thus the choice for the initial standard deviation does not alter the higher order coefficients.

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 interac-

tive 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, we would like to include the duration of the utterance in the interactive application as well.

One 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.

7 Acknowledgements

This work has been funded by the European Social Fund, project POSDRU 6/1.5/S/5 and the national project TE 252 financed by CNCSIS-UEFISCSU.

References

[1] D'Este, F., Bakker, E.: Articulatory Speech Synthesis with Parallel Multi-Objective Genetic Algorithms. In: Proc. ASCI (2010)
[2] Fujisaki, H., Ohno, S.: The use of a generative model of F0 contours for multilingual speech synthesis. In: ICSLP-1998, pp. 714–717 (1998)
[3] Fukumoto, M.: Interactive Evolutionary Computation Utilizing Subjective Evaluation and Physiological Information as Evaluation Value. In: Systems Man and Cybernetics, pp. 2874 – 2879 (2010)
[4] Hansen, N.: The CMA evolution strategy: A tutorial. Tech. rep., TU Berlin, ETH Zurich (2005)
[5] Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. In: Proceedings of IEEE International Conference on Evolutionary Computation, pp. 312 –317 (1996)
[6] Holland, H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975)
[7] Latorre, J., Akamine, M.: Multilevel Parametric-Base F0 Model for Speech Synthesis. In: Proc. Interspeech (2008)
[8] Lv, S., Wang, S., Wang, X.: Emotional speech synthesis by XML file using interactive genetic algorithms. In: GEC Summit, pp. 907–910 (2009)
[9] Marques, V.M., Reis, C., Machado, J.A.T.: Interactive Evolutionary Computation in Music. In: Systems Man and Cybernetics, pp. 3501–3507 (2010)
[10] McDermott, J., O'Neill, M., Griffith, N.J.L.: Interactive EC control of synthesized timbre. Evolutionary Computation **18**, 277–303 (2010)
[11] Moisa, T., Ontanu, D., Dediu, A.: Speech synthesis using neural networks trained by an evolutionary algorithm. In: Computational Science - ICCS 2001,

Lecture Notes in Computer Science, vol. 2074, pp. 419–428. Springer Berlin / Heidelberg (2001)

[12] Panait, L., Luke, S.: A comparison of two competitive fitness functions. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '02, pp. 503–511 (2002)

[13] Qian, Y., Wu, Z., Soong, F.: Improved Prosody Generation by Maximizing Joint Likelihood of State and Longer Units. In: Proc. ICASSP (2009)

[14] Sakai, S.: Additive modelling of English F0 contour for Speech Synthesis. In: Proc. ICASSP (2005)

[15] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: A standard for labeling English prosody. In: ICSLP-1992, vol. 2, pp. 867–870 (1992)

[16] Stan, A., Yamagishi, J., King, S., Aylett, M.: The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Communication **53**(3), 442 – 450 (2011). DOI DOI:10.1016/j.specom.2010.12.002

[17] Tao, J., Kang, Y., Li, A.: Prosody conversion from neutral speech to emotional speech. IEEE Trans. on Audio Speech and Language Processing **14**(4), 1145–1154 (2006). DOI {10.1109/TASL.2006.876113}

[18] Taylor, P.: The tilt intonation model. In: ICSLP-1998, pp. 1383–1386 (1998)

[19] Teutenberg, J., Wilson, C., Riddle, P.: Modelling and Synthesising F0 Contours with the Discrete Cosine Transform. In: Proc. ICASSP (2008)

[20] Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. IEICE - Trans. Inf. Syst. **E88-D**, 502–509 (2005)

[21] Zen, H., Nose, T., Yamagishi, J., Sako, S., Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0. In: Proc. of Sixth ISCA Workshop on Speech Synthesis, pp. 294–299 (2007)