The dramatic piece reader for the blind and visually impaired

Milan Rusko[1], *Marian Trnka*[1], *Sakhia Darjaa*[1], *Juraj Hamar*[2]

[1] Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia
[2] Department of Aesthetics, Comenius University, Bratislava, Slovakia
milan.rusko@savba.sk,trnka@savba.sk, utrrsach@savba.sk, juraj.hamar@sluk.sk

Abstract

The paper presents the concept and realization of the intelligent audio-book reader for the visually impaired. The system is capable of presenting personalities of different characters. The synthesizer mimics the way how a puppeteer portrays different characters. A traditional puppeteer generally uses up to a dozen different marionettes in one piece. Each of them impersonates a character with its own typical voice manifestation. We studied the techniques the puppeteer uses to change his voice and the acoustical correlates of these changes. The results are used to predict appropriate settings of the parameters of the voice for every character of the piece. The information on the personality features of every particular character is inserted manually by a human operator. Similarly to the puppeteer's show only one speaker's voice is used in this concept and all the modifications are made using speech synthesis methods.

Index Terms: audio-book, speech synthesis, personality

1. Audio-books for the blind in Slovakia

The first audio-book in Slovakia was published 50 years ago. The first studio specialized to audio-book recording was founded in 1962 and the first four audio-books were published. The Slovak Library for the Blind – SLB (Slovenská knižnica pre nevidiacich Mateja Hrebendu v Levoči) has now about 37 thousands of library units, from which about 5 000 are audio-books. These books have been read by professional actors - readers. Some of these readers are working for SLB for more than 30 years and some have recorded more than 865 titles [1].

2. Text-to-speech reading eBooks

Reading the book by an actor is time consuming and costly. It takes about two weeks for a professional audio-book narrator to record a novel of a length of about 85.000 words. To fully produce it takes another 2-3 weeks. [2] Moreover the actors and recording studio are not always available. Therefore, the authors of this paper started to cooperate with the SLB library in order to make much more books available – via reading by advanced expressive speech synthesis system.

Text-to-speech (TTS) uses speech synthesizer to read out the given text. TTS in English and some other languages is built into the Windows and Macintosh computer operating systems, phones, tablets and other devices. (The choice of synthetic voices for Slovak was very limited until recently, and there was practically only one producer of professional quality Slovak speech synthesizers in Slovakia – The Institute of Informatics of the Slovak Academy of Sciences.)

The main advantages of eBooks with text to speech over performed audio books is the availability, ease of access and new titles becoming available much quicker. [3]

Several authors have checked the possibilities of expressive speech synthesis for storytelling (e.g. [4] [5]). So did the authors in this study, but their aim was to design a system capable of creating a unique voice for each character.

The Slovak Library for the Blind has made first two synthesized audio-books available for informal evaluation on their web site and presented them also on the international conference Accessibility of audiovisual works to visually impaired people.[6] Two versions were published - one synthesized by unit selection synthesizer Kempelen 2.1 [7] and the second one by statistical parametric synthesizer Kempelen 3.0 [8]. As it was referred in [6] it can be seen from the e-mail reactions of the visually impaired that the quality of both synthesizers was assessed as acceptable with a slight favoring of the unit selection synthesizer. This one was rated as a voice that sometimes sounds almost indistinguishable from human.

The problem with synthesized speech is that it has smaller variability than the natural speech and it becomes tedious after a short while. Therefore the authors decided to prepare and verify a new concept of semi-automatic synthetic audio-books generation, a concept that they called DRAPER - the virtual dramatic piece reader. The idea is that the synthetic or virtual reader should not only read the text of the dramatic piece, but that it should change its voice according to the character being depicted. This concept stems from the former research on the presentation of the personality of the dramatic characters by puppeteers.[9][10] The authors think that the approach of deriving all the synthetic voices from the original voice of one voice-talent has an advance to fulfill the requirement of consistency of chosen voices for audio-book reading, that: " … the voice for each character has to not only be distinctive and appropriate for the character in isolation, but it must also make sense in an ensemble of various characters." [11].

A traditional Slovak puppeteer generally used up to a dozen different marionettes in one piece. Each of them impersonated a character with its own typical voice manifestation. The authors have therefore studied the techniques used by the puppeteers to change their voice and the acoustical correlates of these changes. The prototypical characters (archetypes) were identified. The psychological and aesthetic aspects of their personalities were studied and acoustic-phonetic means of their vocal presentation by the actor were identified [10].

3. Speech synthesizers

The modern speech synthesis system development is dependent on speech databases that serve as a source of synthesis units or a source of data needed to train the models. In the current work we use some of our earlier results, such as neutral speech database [12] and unit-selection synthesizer [7]. On the other hand the expressive databases and expressive HTK voices belong to the most recent results of our research.

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 83–91,
Grenoble, France, 21–22 August, 2013.

3.1. Speech databases

The set of speech databases containing the voice of our voice talent consists of:

1. Neutral database (Level 0) – 2000 sentences
2. Expressive speech database with higher levels of voice effort
    a. Base level (Level 1) – 300 sentences
    b. Increased level (Level 2) -300 sentences
    c. Highly increased level (Level 3) - 300 sentences
3. Expressive speech database with lower levels of vocal effort
    a. Base level (Level -1) - 150 sentences
    b. decreased level (Level -2) 150 sentences
    c. Highly decreased level (Level -3) 150 sentences
4. Whispered speech database - 150 sentences

The Neutral database, VoiceDat-SK, serves for creating the neutral voice with good coverage of synthesis elements.

The method of development of smaller expressive databases that serve for adaptation to voices with higher and lower expressive load (limited to the dimension of emphasis and insistence) was published in [13]. One of the features that are known to be correlated with the level of arousal and vocal effort is the average F0. Figure 1 shows the histograms of F0 for our three databases with one reference-neutral and two increased levels of vocal effort. Histograms of F0 for the expressive databases with one reference-neutral and two lower levels of arousal are presented in Figure 2.

A Gaussian approximation is added to each of the histograms.

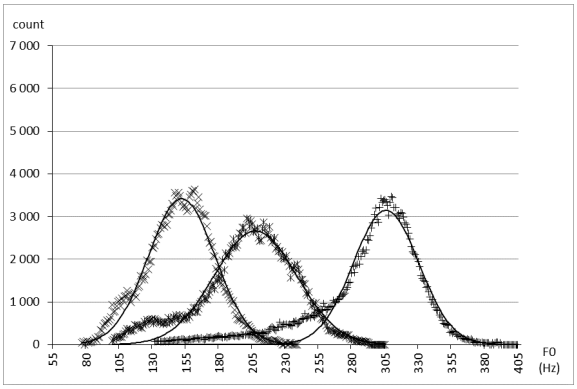

Figure 1: *Histograms of F0 for the three databases with increased vocal effort (from left to right: Level 1, 2, 3).*

In the databases with increasing expressive load the second and third levels of expressivity are clearly distinguishable from the base (reference) level 1. In addition to the neutral voice it is therefore possible to train two more significantly different expressive voices - one with higher and the second one with very high emphasis and insistence.

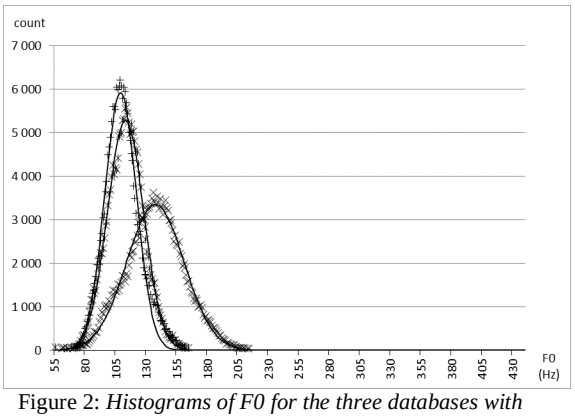

Figure 2: *Histograms of F0 for the three databases with decreased vocal effort (from left to right: Level -3, -2, -1).*

In the databases with decreasing expressive load it was very hard for the speaker to make the second and third levels distinguishable one from another. The differences in vocal intensity and timbre were small and the average F0 was nearly the same for these two databases (level -2 and -3 of expressive load – soothing and very soothing speech). This was probably due to a physiological limit - the lowest frequency of oscillation of the glottal chords. We therefore decided to train only one voice with low expressive load. So we at last came to the choice of speech modes which is identical to the modes examined by Zhang and Hansen from the point of view of vocal effort in their work on classification of speech modes [14] (i.e.: whispered, soft, neutral, loud and shouted in Zhang's description).

A special database of whispered voice was created by the same speaker whispering the same set of 150 phonetically rich sentences as was used in the preceding expressive databases. As it turned out this volume was sufficient to achieve a good quality of synthesized whisper by direct training the HMM voice on this database, without using the neutral voice and adaptation. This is probably due to the absence of voiced parts, which are critical in HMM synthesis because of the problems with pitch tracking. In contrast to voiced parts of the other HMM voices the vocoder buzz is nearly unobservable in the synthesized whisper.

3.2. Synthesizer voices

The authors have several types of synthesizers available derived from the voice of the same voice talent [15]. Two of the used synthesis methods provide sufficient quality for the audio-books reading – the Unit-selection [16] and Statistical-parametric synthesis [17].

Kempelen 2.0 unit-selection synthesizer utilizes the Neutral database with a CART [18] [19] prosody model consisting of F0 model and segmental lengths model. It offers possibilities to change average F0 (AvgF0), to linearly change average speech rate (AvgSr) and to change the depth of application of the prosody model (PDepth). Figure 3 shows the interface for setting these parameters and checking the resulting voice.

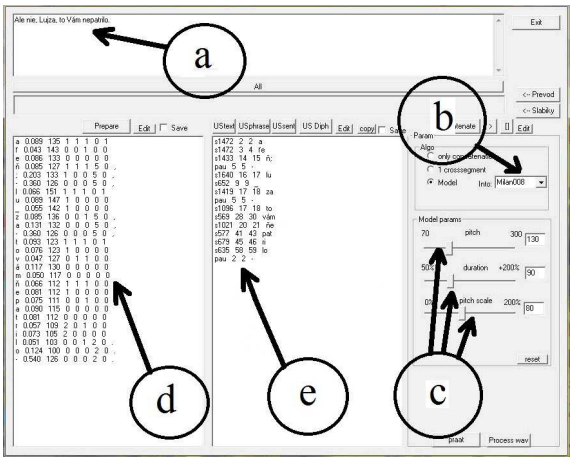

Figure 3: *The graphical interface of the unit selection synthesizer: a) text, b) prosody model selection, c) sliders for setting the AvgF0, AvgSr and PDepth, d) needed phonemes e) best syllable candidates found in the database from which the utterance is concatenated.*

Only neutral voice Unit-selection is available in the actual version of DRAPER as the volume of expressive databases is too small to create good quality expressive Unit-selection voices from them. However certain changes in expression and fitting the voice to different characters can be obtained by changes in average pitch, average speech rate and the weight of prosody model, which influences the depth of the prosody modulation. The last one can change the intonation from monotonous to exaggerative intonation obtained by extrapolation of the model values up to 200%.

Other voices are based on Statistic-parametric speech synthesis [17]. We created the neutral voice in the HTS [20] system and adopted it to three other voices using smaller expressive databases. The statistical parametric synthesis uses Hidden Markov Modeling, and therefore this method is often denoted as HMM synthesis.

The brand name of our synthesizers is Kempelen. The full set of synthesizer voices available in the current version of DRAPER is the following:

- Kempelen 2.1 neutral Unit-selection voice
- Kempelen 3.0 neutral HMM voice
- Kempelen 3.1 HMM voice with higher expressive load
- Kempelen 3.2 HMM voice with very high expressive load
- Kempelen 3.3 HMM voice with lower expressive load
- Kempelen 3.4 HMM whispering voice

4. Personality and voice

In this chapter we shortly introduce our previous research on the relationship between personality and voice.

4.1. Individuality and its components / the notion of personality

In our work, individuality is understood as a psychological entity, a unit consisting of three components [21]. First, personality, is a rather stable component that remains practically unchanged during the life. Second, mood, may change in time, but its changes are rather slow. Third, emotions, are the most dynamical and can change rapidly. In this paper we focus mainly on the first component, personality, and leave the other two components for subsequent studies.

4.2. Traditional psychological classification of personality dimensions

Personality is most commonly described using the formal psychological model called the Five Factor Model [22], [23] with factors representing the basis of the personality space. We have adopted a very simple annotation convention. Each of the personality dimensions - Neuroticism, Extraversion, Openness to experience, Agreeableness and Conscientiousness - will be assigned only one of three values: 1, 0 or -1. For instance, N1 denotes that the character is neurotic, N0 means that this dimension is not applicable, i.e. not important for expressing the character's personality, and N-1 denotes the absence of neuroticism, i.e. that the character comes across as secure and confident, which is opposite to neurotic (see Table 1.).

4.3. Semantic dimension – taxonomy of characters based on elementary semantic oppositions

Personal characteristic of the theatre characters is a complex of three interconnected levels: semantic, visual and acoustical. The semantic level characterizes the character as to its function in the play. The visual layer represents all components of visual representation of a puppet (face, costume, material and animation). The acoustical layer includes speech, music and all the sounds generated by actor and his puppets.

The description of personalities encoded in the speech and acting of puppet characters requires at least a three dimensional space consisting of semantic, visual and auditory dimensions. In the semantic domain the character is best described following its functions in the play. Table 2 gives classification of functions of the characters (the concept) on the basis of elementary semantic oppositions, proposed by us for aesthetic and semantic description on of characters for the purposes of this study.

The classification includes eight dimensions. Some of them are binary, e.g. Sex or Anthropological view, some are coarsely continuous, e.g. Social Status or Ethnicity, and some represent a fine-grained continuum, e.g. Age, Morality, or Intelligence. We will initially code all dimensions as binary since even Age, Morality, or Intelligence are considered archetypal and thus extremely polar for the purposes of the plays.

Table 1. *Five Factor Model of personality with description and examples*

Personality dimension	**Code value**	**Description**	**High level [1] (example adjectives)**	**Low level [-1] (example adjectives)**
Neuroticism	N 1,0,-1	Tendency to experience negative thoughts	Sensitive Nervous Insecure Emotionally distressed	Secure Confident
Extraversion	E 1,0,-1	Preference for and behaviour in social situations	Outgoing Energetic Talkative Social	Shy With-drawn
Openness to experience	O 1,0,-1	Open minded-ness, interest in culture	Inventive Curious Imaginative Creative Explorative	Cautious Conserva-tive
Agree-ableness	A 1,0,-1	Interactions with others	Friendly Compassio-nate Trusting Cooperative	Competitive
Conscien-tiousness	C 1,0,-1	Organized, persistent in achieving goals	Efficient Methodical Well organized Dutiful	Easy-going Careless

Table 2. C*lassification of the characters based on several elementary semantic oppositions.*

Criteria	**One pole**	**code**	**Second pole**	**code**
Sex	Male	XM	Female	XF
Anthropological view	human	HH	Non-human	HN
Age	Old	AO	Young	AY
Morality	Positive	MP	Negative	MN
Aesthetics	Tragical	ET	Comical	EC
Reflexion of ethnic	Our	RO	Foreign	RF
Intelligence	Clever	IH	Stupid	IL
Social status	Noble	SN	Low	SL

For deeper understanding of the actor's notion of personality of his characters, we asked a puppeteer, Anton Anderle, to characterize the personality and to explain the changes of his voice he uses to present them. The actor based the description on both psychological features of the character and the acoustic-phonetic means to express them.

He presented us a set of archetypical characters and their typical features:

I. NEGATIVE MALE TYPE - Intriguer, bad knight

- High volume, hyper-articulation

II. POSITIVE MALE TYPE -Leading man - Royal type dignified, deliberate, wise - *Low pitch, monotonous*

IV. BAD MAN - hoarse, low pitch

V. SWAGGERER Convivial, bold farmer, folk type, straight man, unshuffling, not cunning, frank - *Pharyngeal resonance, great pitch range*

VI. LEAD WOMAN - young, *soft modal*

VII. OLD WOMAN – *lower voice*

VIII. BAD OLD WOMAN Cunning, sarcastic - *Increased* hoarseness, articulator setting as for smile

IX. GOOD OLD WOMAN - Low falsetto, medium pitch range

This actor's classification scheme in fact assigns personality features and semantic features to the acoustical features of the character's voice.

4.4. Voice quality and settings of the articulators

A common way of describing voice settings uses the notion of a reference or neutral setting. This reference corresponds to a normal position relative to possible adjustments [24]. Discrete or continuous variation in voice settings is then depicted as deviations from the reference/neutral setting.

Following the basic pattern outlined in Laver's work [24], one can then attempt to classify voice qualities primarily in terms of description of the position of the articulators. For annotation we used a simple set of labels derived from Laver's terminology e.g. Labial protrusion = LP, Laryngopharyngealized = LPH, Denasal = DN, Harsh whispery creaky falsetto = HWCF, etc. Laver´s classification scheme is considered to be carefully worked-out, and it is being used widely. Despite this, however, some speech qualities are not covered, e.g. smiling or weepy speech. In producing these types of speech, complex positioning of the articulators (wide high/top and wide and low/bottom mouth corner positioning) along with special phonation modes (vibrato etc.) are used, and these are not included in the scheme. Pathological phenomena occurring in spoken utterances, whether acted or natural, such as lisping, stammering, muttering are not included in the scheme either; we have added the PAT (Pathological) annotation mark for them.

Considering prosodic features, we denote slow speech as SRL (Speech rate low), fast speech as SRH (Speech rate high), large pitch range as PRH (Pitch range high), small pitch range as PRL (Pitch range low), and low voice pitch is denoted as LOW.

A complex feature covering both voice quality and prosody is vocal effort. We denote high vocal effort as VEH (Vocal effort high) and low vocal effort as VEL (Vocal effort low).

4.5. Relationship between personalities and acoustic characteristics

We have analyzed 24 voices (the actor's own voice and 23 characters) presented by a puppeteer and we summarize the results in Tables 3 and 4. The numbers representing the highest observed correlation are written in Bold and highlighted. These data can be used for a first analysis of mutual links among personality factors, semantic and articulatory-acoustic features.

As expected, the 2D analysis performed on a relatively limited number of data – does not provide clear answers to the queries related to coding of personality characters by aesthetic-semantic and acoustic speech means. However, the results in Table 3 still suggest some dependencies. For example, negative moral features (MN) can be observed with neurotic (N1), extrovert (E1) and competitive (A-1) characters. Comical characters (EC) are often neurotic (N1). High social position (SH) is connected with calmness (N-1), extroversion

(E), openness to new impressions (O) and strong-mindedness (C). Similar personality characteristics also tend to correlate with wisdom (IH). Results in Table 4 suggest that actors use mostly pitch changes in their voice (LOW+F+CF=22.95%) to express diversity of characters. While female voices (F+CF=12.57% of the total of assigned acoustic marks) are naturally expressed by falsetto, low voices (LOW=10.38% acoustic marks) correlate robustly with the N-1 factor, i.e. with calm and self-assured nature, and obviously with orderliness and resolution (C1).

Additionally, most often used acoustic means include speech rate (SRH+SRL=12.57%) and voice effort intensity (VEH+VEL=12.57%). High speech rate is usually related to neuroticism (N1), extroversion (E1), but also to competitiveness and assertiveness (A-1). On the other hand, slow speech (SRL) tends to be linked to reliability (C1). Considerable range of frequencies of the basic tone in melodic structures (PRH) and high voice effort (VEH), have also been used several times to express neurotic and extrovert nature. More data would be necessary for us to be able to evaluate the function of additional voice properties.

5. Texts of dramatic pieces

DRAPER is meant for reading pieces from various areas of dramatic art in future. However it is still under development and it was decided to prove the concept first on the set of traditional puppet plays. Therefore we use for our first experiments a collection of the texts of puppet shows covering most of the repertoire of the traditional folk Slovak puppeteer Bohuslav Anderle (father of Anton Anderle who presented the puppeteer art to us). The pieces were recorded by Bohuslav Anderle himself in nineteen-seventies and reconstructed, transcribed, edited and published recently by one of the authors of this study, Juraj Hamar [21]. The collection consists of 28 complete puppet plays.

One could reasonably argue that there is no need to create synthesized versions of the games if there are recordings of the text spoken by the puppeteer. The sound quality of the original recordings is very low and is therefore not suitable for publication. On the other hand it can serve as a good study material and reference in evaluation of the quality of our first synthesized dramatizations.

6. Dramatic Piece Reader DRAPER

We have developed a software system for reading texts of dramatic works of art and called it “Dramatic Piece Reader - DRAPER”. It makes use of available set of synthesizers with different expressive load and with wide possibilities to change the characteristics of voices. The schematic diagram of DRAPER is shown in Figure 4.

6.1. DRAPER architecture

With a help of human expert, the Operator, who controls, checks and fine-tunes the text pre-processing and voice assignment, the system creates sound-files of dramatic pieces where every character have a special voice with default acoustical characteristics automatically predicted according to the simple Operator's description. Illustrative sounds can be added wherever it is appropriate (see the following chapter).

After the operator has chosen the text of the dramatic piece to be read the automatic text preprocessing is done. It automatically identifies the characters and shows the list of the characters to the operator. For every character the operator has to manually choose the type of every character (see Table 5).

Table 3. *Counts and mutual occurrences of personality dimensions and semantic characteristics.*

***	XM	XF	HH	HN	AO	AY	MP	MN	ET	EC	RO	RF	SH	SL	IH	IL	*SUMA*	%
Neurotic	3	**4**	**4**	2	2	1	2	**4**		**4**	3		1	**4**		3	**37**	**13,41**
Confident	**4**	2	**4**	1	2	1	3	1	1	1		1	3	1	3		**28**	**10,14**
Extravert	**7**	2	**7**	1	3		2	4		3	2	2	2	5		2	**42**	**15,22**
With-drawn	2	1	**3**		1		1	1		2	1	1		**3**		1	**17**	**6,159**
Open	**3**	2	**3**		1	1	2		1	1	1		2	1	**3**		**21**	**7,609**
Conservative	2	2	**4**		1	1	3	1		2	1	1		3			**21**	**7,609**
Agreeable	**5**	4	**5**	3	2	1	**5**	2		2	1		2	3	3	2	**40**	**14,49**
Competitive	2	3	3	2	2			**5**	1				2			1	**21**	**7,609**
Conscientious	**8**	3	6	4	2		4	3	1	1		1	4	1	2		**40**	**14,49**
Careless	1		**2**		1					1		**2**		**2**			**9**	**3,261**
SUM	**37**	**23**	**41**	**13**	**17**	**5**	**22**	**21**	**4**	**17**	**9**	**8**	**16**	**23**	**11**	**9**	**276**	**100**
%	13,4	8,3	14,9	4,7	6,2	1,8	8,0	7,6	1,4	6,2	3,3	2,9	5,8	8,3	4,0	3,3	100	***

Table 4. *Counts and mutual occurrences of personality dimensions and voice characteristics.*

***	PAT	SRH	SRL	PRH	PRL	VEH	VEL	LOW	WV	F	CF	HV	RL	TV	MV	LV	LS	CR	LP	LL	N	BV	DN	*SUMA*	%
Neurotic	**3**	**4**		2		3				3	1		2	1			2				2			**23**	**12,57**
Confident			3		2		2	**5**	1	2		1				2		2	1	2		1		**24**	**13,11**
Extravert	**2**	**4**		2		3		2		2		1	1	1	1		2				1	2	2	**26**	**14,21**
With-drawn	**1**		**2**		1		1					1									1			**7**	**3,825**
Open	**1**			1			**2**			**2**			1	1	**2**									**10**	**5,464**
Conservative		1	**2**		1		1	**2**		1		1	1			1		1		1				**13**	**7,104**
Agreeable	**2**	1		1		2	**3**	2		**5**				1	1	1	2	1	1		2			**25**	**13,66**
Competitive		**3**		1		**3**		1		2	1	1					2			1				**15**	**8,197**
Conscientious			3	1	2	2	1	**5**	1	3	1	2		1	1	2	2	2	1	3		1		**34**	**18,58**
Careless								**2**														**2**	**2**	**6**	**3,279**
SUM	**9**	**13**	**10**	**8**	**6**	**13**	**10**	**19**	**2**	**20**	**3**	**7**	**5**	**5**	**5**	**6**	**10**	**6**	**3**	**7**	**6**	**6**	**4**	**183**	**100**
%	4,9	7,1	5,5	4,4	3,3	7,1	5,5	10,4	1,1	10,9	1,6	3,8	2,7	2,7	2,7	3,3	5,5	3,3	1,6	3,8	3,3	3,3	2,2	100	***

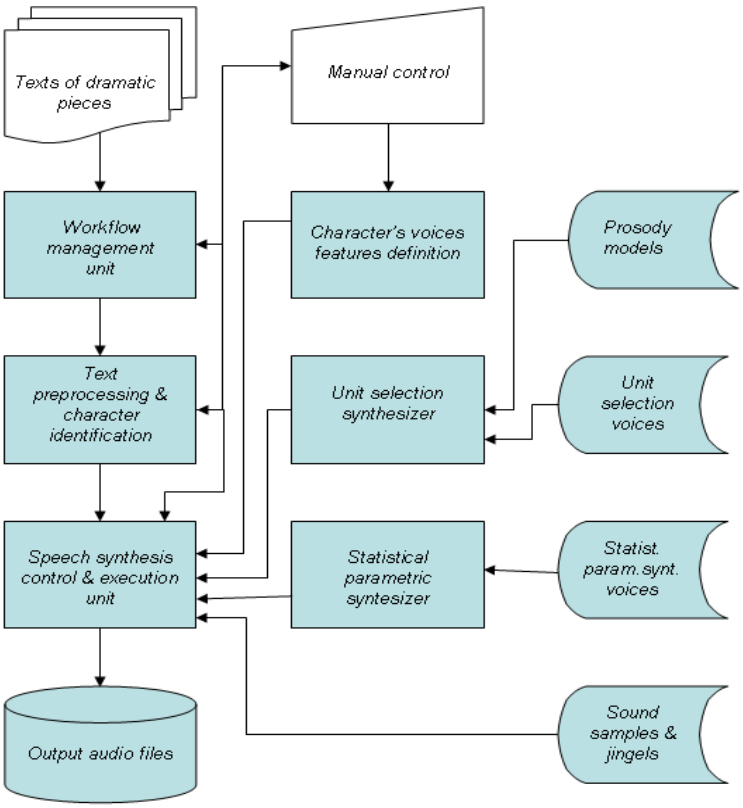

Figure 4: *Schematic diagram of the virtual Dramatic Piece Reader.*

The system then offers a default voice for every character. One arrow means a shift approximately 30 Hz in AvgF0, 15% in AvgSr and 30% in PDepth. The operator can then open the Synthesizer settings window, try the voice with the offered parameter setting and eventually fine-tune the settings for every particular voice. Fine-tuning of the voice can be done also for a smaller part of the text (e.g. whispering or crying only several words).

The sentences ending with exclamation mark are automatically labeled with a tag causing they will be read with expressivity increased by one level (if available).

One of the most important tasks of the operator is to check the text for the words with unusual pronunciation (e.g. names or foreign words) and use the find-replace function of the text editor to replace their written form for the text representation of their pronunciation in Slovak ("write it the way you hear it").

All the preparation of the text for synthesis is technically done by adding tags to the source text. The basic text to be processed is the pure text of the utterances of the characters and the text of director notes, descriptions of the stage and other comments.

Special tags designating the changes of voices and voice parameters are automatically inserted at the appropriate places in the text under the control of Operator.

6.2. A comment on speech temporal dynamics

At various emotions, the same character can have different dynamics of speech. Sometimes he speaks quickly, sometimes slowly, but most often he speaks in a quasi neutral manner. This speed can be set in the Synthesizer settings.

But this is only a so called linear (or horizontal) dynamics applied to the utterance of one character. It is followed (in a vertical direction downwards in the text) by an utterance of another character. The vertical dynamics reflects the dynamics of the dialogue, dynamics of taking the ground, of the speed of changing the characters. If they are speaking calmly, there is a "normal" more or less short pause between the utterances of the characters. However, if the dialogue is expressive (argument, hassle, fear, threats, etc..) the cadence of rotation of the characters in the dialogue is much faster, almost without pause.

Table 5. *The basic set of default voices that the Operator has available.*

Character archetype	Characteristics	Synt. Voice	Avg F0	AvgSr	Pros. Depth
Neutral male	father, comment reader	Unit-sel.	135 Hz = male refe-rence	standard	100%
Leading man	royal type	Unit-sel.	↓	↓	↓
coy	weak	HMM level-2	↑	↓	↓
bigmouth	convivial, folksy	HMM level+2	↑	↑	↑
Negative male	intriguer	HMM level+2	-	-	↑
Very bad man	malicious	HMM level+2	↓↓	↓	↓
Neutral female	mother, comment reader	Unit-sel.	240 Hz = female Refe-rence	standard	100%
Leading woman	royal type	Unit-sel.	↓	↓	↓
Timid woman	shy, un-experienced	HMM level-2	↑	↓	↓
Jovial woman	convivial, folksy	HMM level+2	↑	↑	↑
Negative female	intriguer	HMM level+2	-	-	↑
Very bad woman	malicious	HMM level+2	↓↓	↓	↓
Ghost	Whispering	HMM whisper	-	-	-
Comment reader	neutral to less expressive	HMM level 0	-	-	-

Special marks can be inserted in the text in DRAPER to shorten or lengthen the pauses between replicas of two characters.

7. Commentary reading and Illustrative sounds - Acousticons

One voice has to be dedicated to the Commentary reader, which reads the comments of the theater script, e.g.: “He arrives on the scene; She hits the robber with a stick; He falls to the ground; She catches Kuranto's hand; She hides behind the statue; There is a small cabin near the forest.; etc ...”. This voice should be neutral or with a slightly lower expressivity, but distinguishable from the voices of the characters.

Some actions and phenomena mentioned in the text, for example knocking, ringing, strike, whistle, snoring etc... could be expressed in the acoustic form. Similarly there are different changes of voice qualities, mood, presence of emotions, or speaker non-lexical sounds identified in the text, e.g.: “tearfully, for himself, in a whisper, seriously, screams, angry ...”. Finally, from time to time it is marked in the comment or it is obvious from the text itself, that the character sings the text as a song. Other situations, where the insertion of sounds can be suitable are interjections “Br, brrr” that are usually used when the devil comes. This is also often accompanied by a sound of thundering.

To get an idea of what sounds and what kind of emotionally colored voices are required by the comments, we have analyzed several hundred pages of scenarios of puppet plays.
The examples are from the collection of all 28 games. We list the sounds, emotions or voice modulations that we found in [25].

Sounds: knocking on the door (15 times) , ringing the bell (5 times), whizzing (2 times), striking clocks (4 times), whistling (6 times), snoring (2 times).

Voices: angry (2 times), shouting (4 times), parodying (11 times), crying (13 times), moaning, sobbing (15 times), to himself (13 times), whispering (9 times).

We have therefore included a possibility in DRAPER to insert illustrative sounds in the synthesized speech.

The 256 sounds (acoustic emoticons) are organized in a system of 16 thematically oriented subsets (Transporticons, Zooticons, Sporticons, Eroticons, Partycons, etc.) and are inserted using an easy to remember code. This set of sounds, called SOUNDI we have developed earlier for SMS to Voice service in telecommunications [26].

The Operator can decide which of the instructions (comments) should be read and which should be performed. Some of them can be done by changing the settings of voices and some by insertion of the illustrative sounds.

The letter in the code designates the class and every sound in the class has its own number.

The second way of inserting the sounds is to remember the names of the sound file, which is listed in the full definition of SOUNDI specification (e.g. kiss1 = E1, or gallop = S2).

In further versions of DRAPER the SOUNDI sound database will be enriched and changed substantially including the possibility to use user defined sound samples.

8. Conclusions and future work

Expectations that speech synthesis will be widely used for reading text aloud by readers of electronic books failed to become truth. The reason is that the readers have greater experience from their own reading than listening to synthetic speech, which is often unnatural and is unable to credibly convey the personality of the characters, their moods and emotions.

The possibilities of visually impaired readers are more limited. If the book is not available in Braille, or if their computer is not equipped with Braille display, they would probably like to use to the audio-books. Unfortunately, these are produced in quite a small amount. For this group of people we offer speech synthesis software, which is capable of presenting various characters and their personality.

Similar activities of other researchers in this area [27] [28] indicate that this is a well-grounded approach that will hopefully bring even better effectiveness in producing naturally sounding audio-books in future.

One of the goals of our research was to improve our understanding of the acoustic and auditory correlates of personality dimensions. We introduced a novel approach to the analysis of functional variation, i.e. the need to express personalities of particular characters, in the speech and vocal features of a puppeteer.

Table 6. *The descriprtion of SOUNDI database of sound samples.*

Code	Class	Description	Examples
A1-A16	Acoustic emoticons	sounds reflecting human feelings, moods and attitude to the text	short giggling, laughter, devil laughter, Oooops..., Wow!, Yeees!, sad groan … Sounds suitable for acoustic interpretation of the graphical emoticons.
B1 - B16	Babycons	Acoustic displays of children	children giggling, cry, etc.
E1 - E16	Eroticons	Sounds of love, passion, sex, yearning	kisses, hard beating, sniff, screams, orgasm etc.
V1 - V16	Vulgaricons	Indecent sounds, "dirty sounds" or sounds on the boundary of social acceptability	Fart, belch, spittle, vomit, squelch, hiccup, snore… Whether You like it or not, these sounds belong to the most marketable.
Z1 - Z16	Zooicons	Acoustic displays of animals	Roaster, dog, cat, horse, lion, hen, pig, goat, donkey, mouse, snake, gadfly…
I1 - I16	Symbolicons	Illustrative and symbolical sounds	Church bell, clocks, gun shot, circular saw, glass crack, doors, toilet, etc.
T1 - T16	Transporticons	Sounds of transport means and vehicles	Human steps, horse gallop, car alarm, car crash, car brakes, locomotive, firemen car, ambulance, etc.
P1 - P16	Partycons	Sounds of party and having fun with friends	Filling a glass with a drink, pinging with glasses, opening a bottle of vine, opening a bottle of champagne, sipping, step dancing, Cheers, drunk singing, etc.
S1 - S16	Sportikons	Sports	Table tennis, tennis, judge's whistle, gong, mountaineer falling from a rock, stadium atmosphere, etc.
J1 - J16	Instrumenticons	Jingles or sounds played by musical instruments	Jaw harp, cymbal, church organ, drums, ethnic instruments, etc.
M1-M16	Melodicons	Fragments of the well known melodies with a symbolical meaning	Jingle bells, Happy birthday, Wedding march, etc.

We argued that the system of stylized personality expressions by a puppeteer provides an excellent source of information both for understanding cognitive aspects of social communicative signals in human-human interactions as well as for utilization of observed patterns of human behavior in applications based on interactive voice systems in human machine interactions.

Most important feature of the DRAPER system is, that with a help of human operator it can convert high volume of books into audio form and make them accessible to the blind. The sound-files that will be distributed by the Slovak library for the blind in a form of copy protected files without a violation of the copyright law.

We presented our virtual dramatic piece reader at the conference Accessibility of audiovisual works to the visually impaired - a means of social inclusion and awareness, Bratislava 2012, organized by Blind and Partially Sighted Union of Slovakia. The quality of the generated speech was evaluated as a surprisingly good and acceptable also for longer texts.

At present DRAPER is still under development, but it is already capable of generating sound-files. The formal subjective evaluation tests have not been carried out yet, as we want to further improve our HMM voices through improvements in the vocoder. More work is still needed to make the system less dependent on human operator and to match the automatic text preprocessing to the requirements of this special task.

Our further work will be aimed at adapting the Manual control interface so that it can be operated by a blind person. We also plan experiments with the development of over-articulated highly expressive voice, as the intelligibility of the highest level of expressive speech synthesis is often a bit lower than needed.

Regardless of how 'natural' text to speech can sound, it does not compare to the emotion and performance that an actor can bring to a performed audio book. [3] However the authors of this work try to take steps towards automatic reading of dramatic works in a quality acceptable for the blind and partially sighted people.

Demo sound-files generated by DRAPER can be downloaded from http://speech.savba.sk /DRAPER.

9. Acknowledgements

This publication is the result of the project implementation: Technology research for the management of business processes in heterogeneous distributed systems in real time with the support of multimodal communication, RPKOM, ITMS 26240220064 supported by the Research & Development Operational Programme funded by the ERDF.

10. References

[1] http://unss.sk/sk/aktuality/2012-zvukova-kniha.php, accessed on 19 March 2013.

[2] http://thewritersguidetoepublishing.com/how-does-audio-book-narration-work-heres-the-scoop-from-d-d-scotts-a-mazing-narrator-christine-padovan

[3] http://www.rnib.org.uk/livingwithsightloss/reading/how/ebooks/accessibility/Pages/text-to-speech.aspx accessed on 19 March 2013.

[4] Raimundo, G., Cabral, J., Melo, C., Oliveira, L. C., Paiva, A., Trancoso, I.: Telling Stories with a Synthetic Character: Understanding Inter-modalities Relations, In: Verbal and Nonverbal Communication Behaviours Lecture Notes in Computer Science Volume 4775, 2007, pp 310-323.

[5] Buurman H.A.: Virtual Storytelling: Emotions for the narrator, Master's thesis, University of Twente, August 2007, 113 pages.

[6] Vegh, N.: Commented movies and audio book first realized artificial voice on the website SKN, Accessibility of audiovisual works of art to the visually impaired - a means of social inclusion and awareness organized by Blind and Partially Sighted Union of Slovakia, Bratislava 2012.

[7] Darjaa, S., Trnka, M.: Corpus Based Synthesis in Slovak with Simplified Unit Cost and Concatenation Cost Computation. Proceedings of the 33rd International Acoustical Conference - EAA Symposium ACOUSTICS, High Tatras 2006, Štrbské Pleso, Slovakia. ISBN 80-228-1673-6, pp. 316-319.

[8] Darjaa, S., Trnka, M., Cerňak, M., Rusko, M., Sabo, R., Hluchý, L.: HMM speech synthesizer in Slovak. In GCCP 2011 : 7th International Workshop on Grid Computing for Complex Problems. - Bratislava : Institute of Informatics SAS, 2011, p. 212-221.

[9] Rusko, M., Hamar, J.: Character Identity Expression in Vocal Performance of Traditional Puppeteers. In: Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic. LNAI 4188, pp. 509-516.

[10] Rusko, M., Hamar, J., Benus, S.: Acoustic, semantic and personality dimensions in the speech of traditional puppeteers, Proceedings Coginfocom 2012, Košice, Slovakia, 2012, pp.83-88.

[11] Greene, E., Mishra, T., Haffner, P., Conkie, A.: Predicting Character-Appropriate Voices for a TTS-based Storyteller System, INTERSPEECH 2012.

[12] Rusko, M., Daržágín, S., Trnka, M., Cerňak, M.: Slovak Speech Database for Experiments and Application Building in Unit-Selection Speech Synthesis. In: Proceedings of Text, Speech and Dialogue, TSD 2004, Brno, Czech Republic, pp. 457 – 464.

[13] Rusko, M. Darjaa, S., Trnka, M., Cerňak, M.: Expressive speech synthesis database for emergent messages and warnings generation in critical situations. In Language Resources for Public Security Workshop (LRPS 2012) at LREC 2012 Proceedings, Istambul, 2012, p. 50-53.

[14] Zhang, Ch., Hansen , J., H., L.: Analysis and classification of speech mode: whispered through shouted., Interspeech 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 2007, pp. 2289-2292.

[15] Rusko, M., Trnka, M., Daržágín, S.: Three Generations of Speech Synthesis Systems in Slovakia. In: Proceedings of XI International Conference Speech and Computer, SPECOM 2006, Sankt Peterburg, Russia, 2006, pp. 297-302.

[16] Hunt, A.J. Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database, ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996, pp. 373–376.

[17] Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039-1064.

[18] Breiman, Friedman, Stone, Ohlsen: Classification and Regression Trees. Chapman Hall, New York, USA, 1984.

[19] Rusko M., Trnka M., Darjaa S., Kováč R.: Modelling acoustic parameters of prosody in Slovak using Classification and Regression Trees. In: Human Language Technologies as a Challenge for Computer Science and Linguistics - Proceedings. Poznań, Poland, 2007. ISBN 978-83-7177-407-2, pp. 231-235.

[20] http://hts.sp.nitech.ac.jp, accessed on 13 Dec 2012.

[21] Egges, A., Kshirsagar, S., Magnenat-Thalmann, N.: Imparting Individuality to Virtual Humans. First International Workshop on Virtual Reality Rehabilitation, pp. 201-108, 2002.

[22] Digman, J. M., Personality structure: Emergence of the five factor model, Annual Revue of Psychology, 41, pp. 417-440, 1990.

[23] McRae, R.R., John, O.P.: An introduction to the five-factor model and its applications, Journal of Personality 60, pp.175-215, 1992.

[24] Laver, J., The gift of speech, Edinburgh, UK: Edinburgh University Press, 1991.

[25] Hamar, J. , Puppet Plays of the Traditional Puppeteers, (in Slovak) Slovak Center for Traditional Culture, 2010, 655 pages.

[26] Rusko M., Daržagín S., Trnka M.: "Multilinguality, Singing Synthesis, Acoustic Emoticons and Other Extensions of the Slovak Speech Synthesizer for SMS Reading", ICA 2004, Kyoto, Japan, 2004, pp. IV.3345-IV.3348.

[27] Doukhan, D., Rosset S., Rilliard, A., d'Alessandro, Ch., Adda-Decker, M.: Designing French Tale Corpora for Entertaining Text To Speech Synthesis. LREC 2012: 1003-1010.

[28] Székely, É., Kane J., Scherer S., Gobl Ch., Carson-Berndsen J.: Detecting a targeted voice style in an audiobook using voice quality features. ICASSP 2012: 4593-4596.