Preliminary Results of Emotional Speech Synthesis in Polish[†]

Jolanta Bachan* and Barbara Surmanowicz**

*Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland
**Fachrichtung Elektrotechnik, Wilhelm Büchner Hochschule, Darmstadt, Germany
jolabachan@gmail.com
barbara.surmanowicz@gmx.net

ABSTRACT

Work on expressive speech synthesis has started only in the past few years. Since synthetic speech has achieved high naturalness and intelligibility, many researchers have focused on the need of adding some expressiveness to it to improve the human-computer communication. The present paper reports on a preliminary study of expressive speech synthesis in Polish. The authors look at human perception of synthetic angry, happy, impatient, sad and neutral speech. Additionally, the importance of F0 modelling for expressive speech synthesis is confirmed, whereas the phone duration patterns turned out to have smaller effect on expressiveness. Finally, an experiment investigating influence on human perception of emoticons when used instead of common word labels is presented. The speech synthesis system used is MBROLA with the pl1 female voice and seven neutral male micro-voices created especially for the purpose of the present study. The speech stimuli are generated with the automatic close copy speech (ACCS) synthesis procedure in order to assure the closest similarity to the original model recordings of expressive speech.

1. Introduction

Synthetic speech has been a field of interest of many researchers in the last few decades. Since the synthetic speech became easily intelligible, there are attempts to add more naturalness to it, so that the speech synthesis systems can not only produce intelligible speech output, but can also express personal feelings and attitudes while communicating with humans [1]. While text-to-speech (TTS) systems became widely used to produce neutral speech, often called 'emotionless' speech, the next step of the TTS system development is to make the system generate expressive speech which could convey emotional information. To achieve this goal, one way the researchers have taken is to develop concatenative unit-selection expressive speech synthesisers. But the drawback of such synthesisers is that they rely on a large speech corpora composed of smaller speech corpora representing a collection of expressive speech styles which the synthesisers are to produce [2, 3, 4]. This inconvenience drive many researchers to build limited domain speech synthesisers, where the speech database is

[†]The work was done within a project "Technologies for processing and distributing verbal information in internal security systems", no. R00 035 02.

designed for a particular application and where the quality of synthetic speech gets near perfect [5]. Another solution to avoid creating huge corpora is to use HMM-based synthesis systems, but the low quality of the synthesised speech have pushed the researchers to develop hybrid approaches where the unit selection synthesis and the HMM-based synthesis are unified [cf. 6].

Together with the expressive speech synthesis development, ethical and technical questions have raised. For example, is it ethically correct to conduct a recording of emotional speech where, for instance, the subject is being annoyed only for the purpose of the recording and additionally without his knowledge of being recorded? The elicitation of the real emotions while avoiding ethical problems have been solved in [7, 8]. However, such techniques does not easily allow to create highly structured speech corpora, where, for example, the phonetic coverage in a given language is required. Therefore the researchers use portrayed expressive speech for collecting which the actors are asked to imitate different expressive speech styles. But here the technical problem arises: do the phonetic features of the portrayed speech equal those of the real emotional speech? Studies by [9] and [10] show that acted speech is representative of genuine emotions. Finally, a researcher building an expressive speech synthesiser needs to decide on a set of expressive speech styles to be produced by the system. The literature offers sets containing from 2 to 18 emotions, with a 5-emotion basic set most commonly adopted composed of anger, happiness, sadness, fear, disgust [11, 12].

The aim of the present study is to investigate the effectiveness of synthesising expressive speech based on diphone concatenation, where the diphones are extracted from neutral speech. The expressive speech styles to be produced are angry, happy, impatient,[1] sad and neutral speech. The authors look at the importance of the intonation contours and phone duration patterns for expressive speech synthesis. The amplitude modification and spectrum structure were not investigated at this preliminary stage of the study. Moreover, the influence of speech perception and categorisation is investigated when instead of usual word labels for expression categorisation, emoticons are used.

2. Expressive speech corpus

In the study a small corpus of Polish expressive speech was used. The corpus was originally created for perception tests in diagnosis of Central Auditory Processing Disorders (CAPD) [13]. The corpus contained recordings of 7 different short utterances – containing from one syllable to six syllables in a sentence – uttered by a professional Polish male speaker in 5 expressive states: angry, happy, impatient, sad and neutral. The sentences semantically matched each of the imitated expressive state and their expressive realisation could depend on the overall context. The sentences used in the study were: Tak. *(Yes.)*; Nie. *(No.)*; Dobrze. *(Alright.)*; Dzisiaj nie wolno. *(It's not allowed today.)*; W niedzielę wyjeżdżam. *(I am leaving on Sunday.)*; Wszystko zapomniał. *(He forgot everything.)*; Żegnam na zawsze. *(Goodbye forever.)*

[1]By 'impatient', the authors understand loud, angry and aggressively sounding speech which in the Polish tests was labelled as 'krzyk'. The speaker was asked to produce an angry shout.

The recordings were made in a professional recording studio at the 44.1 kHz sampling rate in stereo. The speaker was asked to produce the sentences imitating the five speech styles. The recordings were annotated manually in Praat on the phone level for the purpose of the present study. Additionally, they were converted to mono and downsampled to 16 kHz as that was the requirement for the software used in the study.

3. Speech synthesis method

The aim of the study is to investigate how effective it is to synthesise Polish expressive speech using a diphone concatenation system with neutral synthetic voice. For this purpose, MBROLA speech synthesis is used for which a female synthetic voice, pl1 [14], already exists and for which it is reasonably easy to provide new voices. The speech synthesis method applied is Automatic Close Copy Speech (ACCS) synthesis [15, 16].

Close Copy Speech (CCS) synthesis or resynthesis method "repeats utterances produced by a human speaker with a synthetic voice, while keeping the original prosody" [17]. In this method, "close copy" means that the synthetic speech is as similar as possible to the original human utterance.

In the present context, "copy" means that the input to the synthesis engine for a given utterance is derived directly from a corresponding utterance in the annotated corpus data. The system consists of a recorded speech signal, a method for pitch extraction from the speech signal, and a time-aligned phonemic annotation of the speech signal. The parameters which are copied from the annotated corpus are phone duration (from the annotation) and F0 (from the speech recording).

Automatic Close Copy Speech (ACCS) synthesis with an MBROLA type diphone synthesis is a process of automatically creating pronunciation specification tables (NLP-DSP interfaces, implemented as PHO files), making use of recorded and annotated real utterances, and synthesising the pronunciation specification tables using an appropriate voice (diphone database). The voice may be created from the annotated utterances, or may be an independently created voice. In the present study both cases of the voice are used.

The ACCS procedure therefore emulates the Natural Language Processing front end to a speech synthesis engine. The speech and annotation information are transformed automatically into a pronunciation specification table which, together with a diphone database, constitutes the input to the synthesis engine, which converts the specification table into speech using the diphone database. The acoustic output is a speech file. Figure 1 shows the general architecture of the ACCS synthesis system [15, 16].

3.1. MBROLA micro-voice creation

For the speech synthesis in the current study two voices were used – female and male. The Polish female MBROLA voice pl1 [14] already existed and was available on the MBROLA project website [18, 19]. The male voice was created from the neutral recordings of the expressive speech corpus used in the study.

From the expressive speech corpus, recordings and their annotation files for only the neutral speech style were selected. For each neutral utterance diphones were ex-

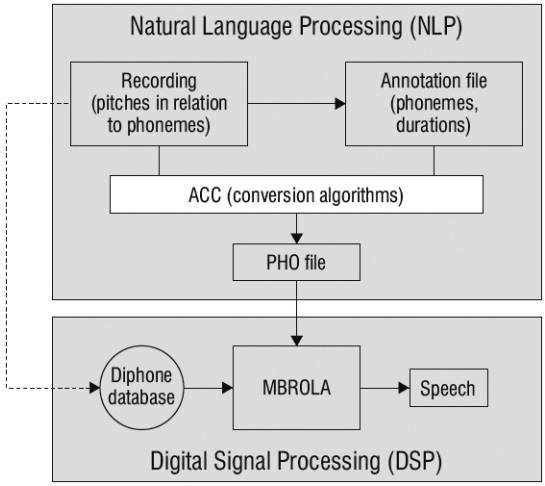

Figure 1. ACCS synthesis system architecture [15, 16].

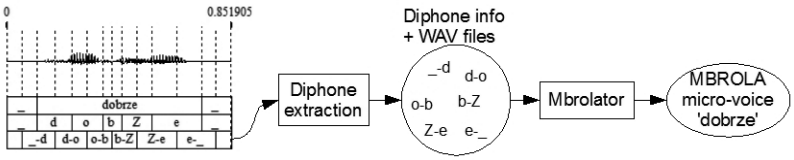

Figure 2. Mbrolation, the MBROLA micro-voice creation procedure.

tracted using a diphone extraction program and then input to the Mbrolator [17, 19], a software for MBROLA voice creation.[2] As a result of this process 7 male neutral MBROLA micro-voices were created, each micro-voice containing a set of diphones corresponding to the diphones in the 7 different sentences from the expressive speech corpus. The MBROLA micro-voice creation procedure, called mbrolation, is shown on Figure 2.

4. Speech perception tests

Three speech perception tests were designed according the EAGLES standards [20]. In the tests 6 native speakers of Polish participated, 3 males and 3 females in the age of 21–28 years old. They took the tests separately in a quiet room and listened to the stimuli via headphones to minimise the ambient noise. The tests were self-paced and the stimuli could be played as many times as the subjects required. Each session lasted 30–90 min.

[2]The authors are grateful to Dafydd Gibbon and Catharine Oertel for the automatic diphone extraction and mbrolation.

4.1. Test 1: CatNum judgement

The CatNum (categorical-numerical) judgement test was designed to examine the general perception of the expressive speech synthesis. In the test three major groups of stimuli were used: (1) original male recordings, (2) automatically close copied original recordings synthesised with the male micro-voices, (3) automatically close copied original recordings synthesised with the Polish female MBROLA pl1 voice. Because of the mismatch between the male pitch and female pitch, the original male pitch values were multiplied by 2 to adjust to the female voice standards (for the effectiveness of this cf. [16, 21]). Altogether 105 stimuli were used: 7 sentences * 5 expressive modes * 3 voices.

The testees were instructed to listen to the recording and to choose one of the expressive labels which fit the recording best. Additionally, the testees were asked to mark on the scale the confidence with which they chose the expressive label. 1 meant little confidence, 5 was the highest confidence level.

The results of the test are shown in Table 1. The table presents a confusion matrix where in the rows there are synthesised expressive styles and the columns show their perception by the test participants. The gray diagonal underlines the perfect match. The *Cat* (category) figures show the total number of scores, i.e. the sum of *Cat* numbers in rows is 42 – meaning 7 sentences * 6 testees. The *Conf* (confidence) figures show the average confidence with which the testees chose the expression label. The figures in bold indicate the scores which were higher than the expected results, i.e. the perfect match.

The results show that in almost all cases the expressive speech styles were recognised correctly. There are only 4 instances (in bold) where the expressive styles were not perceived as expected. Those are:

1. female ‘happy’ voice recognised as ‘impatient’ – the acoustic properties of both speech styles could be misleading;

Table 1. Results of Test 1: CatNum judgement. Cat means category, Conf means confidence

		angry		**happy**		**neutral**		**sad**		**impatient**	
		Cat	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*
Female	angry	11	3,64	3	2,33	9	2,78	8	3,38	11	3,45
	happy	12	3,58	11	3,45	3	2,67	3	2,67	**13**	**3,46**
	neutral	5	3,2	3	2,33	25	3,72	9	3,89	0	–
	sad	2	4	0	–	16	3,38	24	4,08	0	–
	impatient	**20**	**3,35**	3	4	0	–	6	3,5	13	3,23
Male	angry	7	3,29	3	4,33	**20**	**3,3**	9	3,44	3	3
	happy	7	2,57	9	3,56	**14**	**3**	4	3,75	8	2,88
	neutral	4	3,25	0	–	21	4,19	17	4,12	0	–
	sad	1	2	0	–	15	3,47	26	4,15	0	–
	impatient	11	3,45	4	4,25	6	2,67	8	3,63	13	3,54
Original	angry	24	4,17	2	4,5	7	3,43	0	–	9	3,78
male	happy	9	3,67	22	3,95	5	3,2	0	–	6	4
	neutral	3	3,33	0	–	27	4,07	10	3,9	0	–
	sad	2	3	1	4	2	3,5	37	4,78	0	–
	impatient	17	4,82	0	–	0	–	0	–	25	4,92

2. female 'impatient' voice recognised as 'angry' – this can be the result of the imperfect conditions of the original recording and expressive style selection as the 'impatient' was imitated by the speaker as 'very angry loud speech' and ACCS did not provide amplitude modification, so the loudness in the synthesis was not adjusted to the original recording. In the instructions the subjects were asked to mark 'impatient' when they felt that the 'angry' label was not strong enough;
3. male 'angry' and 'happy' voices recognised as 'neutral' – this result may suggest that the male voice is in general perceived as calmer than the female voice, but this needs further investigation.

The perception of the expressive styles of the original voice matches almost perfectly the ideal cases and all the best results are centred on the diagonal, which suggests that the expression imitation of the speaker was correct. However, the high number of the 'impatient' mode marked as 'angry' confirms the imperfection of the expressive speech style selection in the corpus.

The surprising fact is that the expressions of the stimuli synthesised with the female voice had higher scores of correct recognition than the stimuli synthesised with the male voice.

The confidence figures indicate that the least hesitation the subjects had while assessing the original voice and the further from the perfect match on the diagonal, the smaller their confidence.

4.2. Test 2: same/different

In Test 2 pairs of stimuli are compared to see if they are perceived as being the same or different. The test was designed to examine the importance of the F0 contour modelling and the smaller advantage of the phone duration patterns in speech synthesis of the expressive styles chosen for the present study.

In the test one sentence from the corpus was taken and synthesised in different ways using the male MBROLA micro-voice for this sentence. First, this sentence was ACCS synthesised using the neutral speech mode and therefore constituting a group N of stimuli with one item only. Second, the expressive speech modes: angry, happy, sad and impatient were ACCS synthesised creating a group A of stimuli with 4 items. Third, a group B of stimuli was created, in which the stimuli had the following structure: (1) the phone durations were taken from the neutral recording, but the F0 values were taken from the expressive speech recordings, (2) the phone durations were taken from the expressive speech recordings, but the F0 values were taken from the neutral recordings. The group B consisted of 8 stimuli.

The following comparison pairs were created for the purpose of the study: AB / BA / AA / BB / NB/ BN / NN – altogether 51 pairs of stimuli. AA, BB and NN pairs were added as noise to count the false alarm rate.

The test results are presented in Table 2. The results show that when the stimuli with modified F0 and phone duration structure (group B) were compared with the synthesis of the neutral speech style, they were in 90% recognised as being the same, which suggests that the testees did not hear the difference between the neutral speech style and the stimuli with the modified structure. However, the results of the second comparison confirm the original assumption of F0 modelling being more important

Table 2: Results of Test 2: same/different

	neutral	expressive
F0	92,50%	97,50%
durations	90,00%	32,50%

for expressive speech synthesis; when the stimuli from group B with the neutral phone duration values but expressive F0 values were compared with their expressive counterparts, they sounded the same in almost 98% of the cases, whereas stimuli from group B with expressive phone durations and neutral F0 sounded as their expressive counterparts only in 32% of the cases. The latter figure could even been smaller if the results of the 'sad' expressive speech were separated from the rest.

4.3. Test 3: emoticons

The aim of Test 3 was to see the influence of the use of emoticons representing facial expressions of emotional states, instead of usual word labels. For the test five emoticons were chosen corresponding to angry, happy, impatient, sad and neutral facial expressions (see emoticons in Table 3). For the test one sentence was chosen from the corpus and ACCS synthesised using MBROLA female pl1 voice and male MBROLA micro-voice in 5 expressive modes. Additionally, the original recordings of this sentence were added. Altogether, 15 stimuli were used: 1 sentence * 5 expressive modes * 3 voices.

The testees were instructed to listen to the recording and to choose one of the emoticon which fit the recording best. Additionally, the testees were asked to mark on the scale the confidence with which they chose the emoticon. 1 meant little confidence, 5 was the highest confidence level.

The results of Test 3 were compared with the results of the same sentence scored in Test 1, in which the word labels, not the emoticons, were used. The comparison of the results is presented in Table 3.

The table shows a confusion matrix where in the rows there are synthesised expressive styles and the columns present their perception. The gray diagonal underlines the perfect match. The *Cat* (category) figures show the total number of scores. The *Conf* (confidence) figures show the average confidence with which the testees chose the emoticon. The figures in bold indicate the figures which were higher than the expected results, i.e. the perfect match.

The results show that using emoticons instead of word labels introduced more confusion which is depicted by a higher number of figures in bold. Additionally, the confidence was a little bit smaller when the emoticons were used. However, the original male voice and its expressive styles were recognised with less confusion in the test with the emoticons, than in the test with word labels.

5. Conclusions

In the present paper experimental work with expressive speech synthesis in Polish was presented. It was shown that synthesising expressive speech using diphone database

Table 3. Results of Test 3: emoticons, compared with the corresponding values from Test 1. F – female voice, M – male voice, O – original voice; A – angry, H – happy, N – neutral, S – sad, I – impatient; Cat – category, Conf – confidence

		angry				**happy**				**neutral**				**sad**				**impatient**			
		Cat	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*	*Cat*	*Conf*
F	A	2	4,5	5	3,2	0	–	1	1	1	2	0	–	2	4	0	–	1	3	0	–
	H	0	–	0	–	5	3,8	5	3,4	1	4	1	0	0	–	0	–	0	–	0	–
	N	1	4	1	3	**3**	**2,3**	**3**	**2**	2	3	1	4	0	–	1	3	0	–	0	–
	S	0	–	1	3	0	–	0	–	2	3	**4**	**3**	4	4,5	1	4	0	–	0	–
	I	**5**	**3,8**	**3**	**3**	0	–	0	–	0	–	**2**	**2,5**	**1**	**5**	0	–	0	–	1	1
M	A	0	–	1	3	0	–	0	–	**6**	**2,8**	**5**	**3,8**	0	–	0	–	0	–	0	–
	H	1	3	1	4	3	4	4	4	2	4	0	–	0	–	0	–	0	–	1	4
	N	0	–	0	–	0	–	0	–	3	4,3	2	5	3	4,3	**4**	**3,8**	0	–	0	–
	S	0	–	0	–	0	–	0	–	0	–	**5**	**3,6**	6	3,7	1	4	0	–	0	–
	I	**2**	**4**	**2**	**3,5**	0	–	**1**	**4**	**4**	**2,8**	**2**	**2**	0	–	**1**	**4**	0	–	0	–
O	A	4	3,5	5	4,4	0	–	0	–	2	4,5	0	–	0	–	0	–	0	–	1	4
	H	2	4,5	3	3,8	4	4,3	3	4,7	0	–	0	–	0	–	0	–	0	–	0	–
	N	1	4	0	–	0	–	0	–	2	2	5	3,8	**2**	**4**	1	3	0	–	0	–
	S	0	–	0	–	0	–	0	–	0	–	**3**	**3,3**	6	5	3	2,3	0	–	0	–
	I	**4**	**4,8**	1	5	0	–	0	–	0	–	0	–	0	–	0	–	2	5	5	5

extracted from neutral recordings is effective and preliminary perception tests confirmed it. This gives hope that good results of expressive speech synthesis may be achieved using the existing neutral diphone databases, and it may be unnecessary to create new diphone databases for expressive speech styles like angry or happy. This would save a lot of work of speech corpus creators.

The MBROLA speech synthesiser turned out to be a good tool for expressive speech synthesis and manipulation of the speech parameters, such as F0 values and phone durations, but did not allow for amplitude modifications. Additionally, Automatic Close Copy Speech (ACCS) synthesis allowed to create a whole set of different expressive stimuli for the speech perception tests, based on the parameters taken from the original utterances and their annotations, therefore creating the expressive speech synthetic stimuli with the best possible quality for the given synthesiser.

The test with the F0 contour and phone duration modifications showed that for expressive speech synthesis in the selected expressive styles set the intonation contours were more important than the duration patterns.

Finally, the usage of emoticons instead of word labels in expression categorisation introduced more confusion, although the confidence scores were similar to those in the test with word labels. This is quite optimistic if one needs to use pictures instead of word labels, for example in tests with children or work with robots. However, in such cases the set of emoticons or pictures should be chosen carefully to avoid ambiguity.

For future tests it is also necessary to modify the set of the expressive speech styles, as the impatient style, labelled in Polish as 'krzyk' is very controversial and the instructions given to the speaker made 'impatient' and 'angry' styles be different only in loudness, not the actual emotion imitation by the speaker. Since in this study the amplitude modifications were not performed, therefore the 'impatient' speech style should not have been investigated as a separate category, but should be integrated into the 'angry' speech style.

References

[1] Bailly, G. Campbell, N. & Möbius, B. 2003. ISCA special session: hot topics in speech synthesis. In: *EUROSPEECH-2003*, pp. 37–40.

[2] Campbell, N. 2003. Databases of Expressive Speech. In: *Proceeding of the Oriental-COCOSDA Workshop 2003*. 1–3 October 2003, Singapore.

[3] Hunt, A. and Black, A. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: *Proceedings of ICASSP 96*, vol 1, pp. 373–376, Atlanta, Georgia.

[4] Schröder, M. & Grice, M. 2003. Expressing vocal effort in concatenative synthesis. In: *Proceeding of the 15th International Conference of Phonetic Sciences*, pp. 2589–2592. Barcelona, Spain.

[5] Black, A. & Lenzo, K. 2000. *Limited Domain Synthesis*. In: *ICSLP2000*, pp. 411–414. Beijing, China.

[6] Black, A., Zen, H., & Tokuda, K. 2007. Statistical Parametric Synthesis. In: *ICASSP 2007*, Hawaii.

[7] Tolkmitt, F. & Scherer, K.R. 1986. Effect of experimentally induced stress on vocal parameters. In: *Journal of Experimental Psychology: Human Perception and Performance*, 12. pp. 302–313.

[8] Johnstone, T. & Scherer, K.R. 1999. The effects of emotions on voice quality. In: *Proceedings of the XIVth International Congress of Phonetic Sciences*.

[9] Williams, C.E. & Stevens, K.N. 1972. Emotions and speech: some acoustic correlates. In: *Journal Of The Acoustical Society Of America*. v52 i4(2). pp. 1238–1250.

[10] Banse, R. & Scherer, K.R. 1996. Acoustic Profiles in Vocal Emotion Expression. In: *Journal of Personality and Social Psychology*. pp. 614–636.

[11] Ortony, A., & Turner, T. J. 1990. What's basic about basic emotions? *Psychological Review*, 97, pp. 315–331.

[12] Murray, I.R. & Arnott, J.L. 2008. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. In: *Computer Speech and Language*, Volume 22 Issue 2. Academic Press Ltd.

[13] Wojnowski, W., Obrębowski, A., Pruszewicz, A., Demenko, G., Wiskirska-Woźnica, B. & Świdziński, P. 2007. Further research on speech tests (Dalsze badania nad testami mowy utrudnionej). In: *Otolaryngologia polska. The Polish otolaryngology (Otolaryngol Pol)* vol. 61 Issue 6. ISSN: 0030-6657 Poland, pp. 979–82.

[14] Szklanny, K. & Masarek, K. 2002. PL1 – A Polish female voice for the MBROLA synthesizer. Copying the MBROLA Bin and Databases. <http://tcts.fpms.ac.be/synthesis/mbrola/mbrcopybin.html>, accessed 2006-11-25.

[15] Bachan, J. 2007. Automatic Close Copy Speech Synthesis. In: *Speech and Language Technology*. Volume 9/10. Ed. Grażyna Demenko. Poznań: Polish Phonetic Association, pp. 107–121.

[16] Bachan, J. 2007. *Close Copy Speech Synthesis for Perception Testing and Annotation Validation*. M.A. thesis. Adam Mickiewicz University, Poznań, Poland.

[17] Dutoit, T. 1997. *An Introduction To Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

[18] Dutoit, T. 2005. The MBROLA Project. <http://tcts.fpms.ac.be/synthesis/mbrola.html>, accessed 2008-08-16.

[19] Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & van der Vrecken O. 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. In: *Proceedings of ICSLP 96*. Philadelphia, vol. 3, pp. 1393–1396.

[20] Gibbon, D., Moore, R. & Winski, R. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

[21] Gao, Y. & Yang, Z. 2007. Pitch modification based on syllable units for voice morphing system. In: *Proceedings of the 2007 IFIP International Conference on Network and Parallel Computing Workshops*, pp. 135–139.