Linguistic and Acoustic Resources of the Computer System of Spoken English Intonation Training

Zdaranok Y.A.
Belarussian National Technical University
Minsk, Belarus
Email: yuliyazdaranok@gmail.com

***Abstract*—This paper considers supra-segmental parameters as intonation, stress and speech rhythm, which are related not only to the correct articulation of sounds of the target language, but the correct pronunciation depends on prosodic structures presented by a standard intonation patterns. Linguistic and acoustic resources are needed for computer-aided intonation training because they are the basis for pronunciation instructions in and outside the classroom.**

***Keywords*—supra-segmental parameters, intonation, pronunciation, linguistic and acoustic resources, computer pronunciation training**

I. Introduction

Intonation is one of the first aspects of speech, to which the baby reacts, which perceives and makes attempts to reproduce itself. It was revealed that the child has few difficulties in developing a native speaker intonation. Thus, intonation seems more simple and it is taken for granted at the time of amplification of infant speech. It seems that an adult should remember and reproduce intonation much easier but they are making a lot of efforts by learning to intonation structure.

Proper pronunciation of the target language is associated with the correct articulation of sounds and also with supra-segmental parameters. Supra-segmental parameters are aspects of speech, referred to prosody.

By teaching prosody is important to understand and describe the supra-segmental parameters that are detected in the target language. It is also very important to describe the prosodic pattern of speech. The students often have issues producing correct intonation and they must learn to intonate the modalities, otherwise their speech is much less intelligible [1].

Anyway, by learning prosody is necessary to know an intonation contours palette to convey the diversity of thoughts in speech. Therefore, the intonation should be taught in the context of a well-structured dialogue or discourse.

Speech is a universal means of communication. It includes the processes of generation and perception (reception and analysis) messages for communication purposes in all languages of the world where leading thought or mental image is implemented in the speech by acoustic instruments.

The sentence is a combination of grammatically and phonetically structured performance of human thought during the speech. It is known that the sentence possesses definite phonetic features: speech melody, sentence-stress, tempo, rhythm, pauses and timbre. Each feature performs a definite task, and all of them work simultaneously [1]. An utterance consists of one or more phrases. The phrase has a semantic completeness and syntactic structure. The phrase is the largest unit with a complete phonetic intonation. The main distinguished unit in the phrase is the core, around which are concentrated its accompanying elements pre-nucleus and post-nucleus.

II. Intonation pattern of melodic portrait

The present work is a follow up study to the previously introduced model of universal melodic portraits (UMP) of accentual units (AU) for representation of phrase intonations in text -to-speech synthesis [2]. According to this model, a phrase is represented by one or more of AUs. Each unit, in turn, can be composed of one or more phonetic word. If there is more than one word in an AU, than only one word bears the main stress while other words carry a partial stress. Each AU consists of pre-nucleus (all phonemes preceding the main stressed vowel), nucleus (the main stressed vowel) and post-nucleus (all phonemes following the stressed vowel).

The UMP model assumes that topological features of melodic AU for particular type of intonation do not depend on a number or quality of phonemic content of a pre-nucleus, nucleus or post-nucleus, nor on the fundamental frequency range specific for a given speaker. The UMP model allows to represent intonation constructs as a set of melodic patterns in normalized space Time – Frequency.

Time normalization is performed by bringing pre-nucleus, nucleus and post-nucleus elements of AU to standard time lengths. This sort of normalization levels out the differences in melodic contours caused by the number of words and phonemes in an AU.

For fundamental frequency normalization F0 min and F0 max are determined within the ensemble of melodic contours produced by a certain speaker. This sort of normalization cancels out the differences of melodic contours caused by speaker's voice register and diapason.

The normalization is calculated by the formula

$$F_0^N = (F_0 - F_0min)/(F_0max - F_0min). \quad (1)$$

In certain cases, it may be beneficial to use statistical normalization instead of 1

$$F_0^N = (F_0 - M)/\zeta, \quad (2)$$

where M is mathematical expectation, ζ is standard deviation. Note that M can be interpreted as a register and ζ – as a diapason of speaker's voice.

Therefore, the normalized space for UMP may be presented as a rectangle with axes (T_N, F_0^N) as schematically shown in Figure 1, while the interval [0 – 1/3] on the abscess T_N Structure of linguistic resource is a pre-nucleus, [1/3 – 2/3] is a nucleus, and [2/3 – 1] is a post- nucleus. The intervals on the ordinate F_0^N: [0 -1/3] – low level, [1/3 – 2/3] – mid-level, [2/3 – 1] – high level.

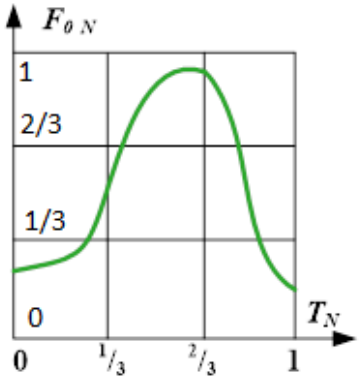

Figure 1. Main view of UMP

III. Structure of Linguistic and Acoustic Resources

In present research we use the resources of English texts and audio-files [3] which included:

- 44 everyday situations, each containing four dialogues in natural conversational English;;
- All dialogues consist 1051 sentences, including 704 affirmative sentences, 325 interrogative sentences and 22 exclamatory sentences, spoken by certain number of male and female speakers;
- Situations relevant to those studying or travelling in England, including eating out, entertainment and travel, as well as more general functions such as greetings, complaining and apologizing.

Each dialogue is structured by:

- the speaker (man and/or woman),
- the number of participants in the dialogue,
- the type of sentence: questions, statements and exclamations,
- the number of phrase units,
- number of AU in the phrase,
- the specification pre-nucleus, nucleus, post-nucleus.

In his "Advice to Foreign Learners" A.C.Gimson emphasizes necessity of learning "the English usage of falls and rises to signify the mood of the speaker, so that an over-use of rises will not give an unintentional impression of, for example, diffidence or complaint, and too many falls create an unwitting effect of impolite assertiveness" [4].

For that reason, the processing of acoustic materials was conducted according to the following intonation criteria:

- the falling tune;
- the rising tune;
- the falling-rising tune.

(a) The falling tune

The voice falls from a high to a low note on one stressed syllable. It is used in short complete statements, for questions beginning with a question word, for question tags when the speaker is sure that what he says is right or for orders and exclamations.

(b) The rising tune

The voice rises on the last stressed word or on the unstressed syllables following the last stress. It is used for statements intended to encourage, for questions which are answered by, for questions beginning with question words when the speaker wishes to show special interest, for question tags when the speaker is not sure that what he says is correct, for sentences ending with "please"; for "goodbye"; for "thank you" when it is used to show gratitude for a simple matter (passing the salt etc.)

(c) The falling-rising tune

The voice falls on the most important part of the sentence and rises again. It is used for apologies, for expressing tentative opinions.

According to the grammar rules of English there are two types of commonly used interrogatory sentences: general and special questions. The statistics below show how often are interrogatory sentences used depending on the situational moment during the interaction.

Statistic: General question

Is	16	Will	6	Need	1	Shall	2
Are	12	Do	26	Has	4	May	2
Am	1	Does	4	Can	23	Must	1
Was	2	Did	2	Could	17	Would	16
Were	1	Have	21	Should	1	TOTAl	140

Statistic: Special question

What	66	Who	2	Whose	0	Whither	0
When	13	How	52	Wherefore	0	Whence	0
Why	2	Which	8	Whatever	0	However	0
Where	14	Whom	0	Wherewith	0	TOTAL	157

Table 1 (English) and Table 2 (Russian) gives us a clear idea that the minimum (F0 min) and maximum (F0 max) of fundamental frequency - F0 differs in the entire ensemble of intonation patterns IKi in the utterance spoken by native English speaker and native Russian speaker. That makes obvious that the voice of pitch in English is higher than the voice of pitch in Russian.

Intonation type	Affirmative		Special		General	
F0 [Hz]	min	max	min	max	min	max
Sample 1	92	184	100	330	109	280
Sample 2	90	180	98	280	98	286
Sample 3	100	230	60	235	101	252
Sample 4	105	230	65	232	99	211
Mean value	96.75	206	80.75	268.25	101.75	257.25
Diapason	2.13		3.34		2.53	

Intonation type	Affirmative		Special		General	
F0 [Hz]	min	max	min	max	min	max
Sample 1	80	147	85	154	85	170
Sample 2	78	150	85	155	91	196
Sample 3	81	144	80	155	84	185
Sample 4	82	146	83	157	84	185
Mean value	80.25	146.75	83.25	157.5	86.25	182.25
Diapason	1.83		1.89		2.11	

Table 1. shows the minimum (F0 min) and maximum (F0 max) of fundamental frequency in English

Table 2. shows the minimum (F0 min) and maximum (F0 max) of fundamental frequency in Russian

IV. The overview of computer-aided language learning programs

Computers were used for language learning since 1960 in the last century [5]. The research in this area of computer-aided language learning can be divided into two equal sphere from technical point of view. The whole system has got three main fields: early systems, new voice input systems and dialogue based systems. On the one side, there are various systems that have a form of websites with fill in the gaps tasks, online chat, statistic multimedia programs, modifications of popular games or even a set of digital music files for playback. On the other side, systems are able to natural language understanding, voice synthesis, and high interactive 3D programs to teach a cultural norm and also a language. For example, there are programs just for vocabulary learning, but some are focused on grammar learning. Computer-aided pronunciation programs can be separated in two very small categories such as those which are used for speech segmentation learning, and those which gives instruction on phrase level or discuss level.

A. Early systems

The Programmed Logic for Automatic Teaching Operations (PLATO) [6] system was one of the early system of computer-aided language learning which was running on the largest and expensive mainframe. PLATO and other similar systems were based on text that was presented to students with task and advise to fill it with appropriate words. If the answers were wrong, the program informed them about that without explanation of the errors. The pejorative monikers were used to describe the monotonous aspect of systems of this type. IBM developed a specialized hardware and programmed material for German language learning in New York university with tasks aimed to fill in the gaps accompanied by pre-recorded audio tracks.

B. Modern systems

Modern systems as usual perform more opportunities for language learning which includes high quality audio, graphics and automated feedback. The content of the lessons is not static and is generated randomly as a response of student action.

Many systems use some forms of automatic system recover, speech synthesis, natural language understanding, or natural language generation. WebGrader [7] was a tool for pronunciation training which helps students of French to obtain an automatic pronunciation evaluation feedback based on on calibrated machine scores. One of interesting results was that the students were disappointed with the scoring, sometimes it seemed to be wrong, but there wasn't opportunity for segmentation of sentences and sending a feedback.

The Voice Interactive Language Training System (VILTS) [8] used a task-based language learning approach. The learning activities were divided into three separated levels of category of activity (speaking, reading and listening). The graphic user interface suggested the way in which the lessons can take place, but students have chosen their own way for language learning. The study showed that students reacted positively to the system, finding, that the navigation is userfriendly, and natural language recognition occurs in interactive activities and pronunciation feedback were all important factors in positive program reception.

The Tactical Language Tutoring System (TLTS) is a good example of a rich multimedia system for language learning. The student is immersed in the 3D word of language using virtual tournament where he is instructed to fulfill the tasks. Speech recognition is performed using the Hidden Markov Model Toolkit and augmented with noisychannel models to capture mispronunciations [9]. The computer-aided language learning system created dynamic questions for practice on base of teacher's sentence pattern. Graphical representation of the parts of the sentences which are practice on a given moment were shown to prompt the student to generate the whole sentence. A grammar network was created to grab potential errors according to the greatest impact where the impact was estimate as a increase of grammar errors divided by the increase of nonagreement in the model.

C. Dialogue based training system

Dialogue based training systems are used to create a virtual environment in which students ca hold dynamic and natural conversations. Instead of a given specific sentence or limited script that can lead the student to memorise them in the learning process, students can hold conversations that are varied between practice sessions. The speech recognition technology is imperfect, there is continuous tension in the system of dialogue between allowing freedom in conversation and sufficient constraining way of acting to be held in the given parameters.

V. Computer-Aided Pronunciation Training

The system of computer-aided pronunciation training (CAPT) is created to evaluate and improve pronunciation in

foreign language. The computer-aided pronunciation training system can be considered as an evaluation component and a feedback component. Pronunciation evaluation can take place at two general levels: holistic and pinpoint errors detection. A holistic evaluation considers a wide choice of speech and gives the whole evaluation of speaker's proficiency. A pinpoint errors detection attempts to detect the concrete mispronunciation on a word and subword level [10].

A. Methods of pronunciation evaluation

There are several methods suggested to holistic evaluation of pronunciation. All of them includes correlation of subjective human measures with based on machine measures. The acoustic and probabilistic measurements include total duration of read speech without any pauses, total duration of speech with pauses, mean segment duration, rate of speech, and log likelihood measurements. The human rating includes the quality of pronunciation, segmentation quality, fluency and rate of speech.

B. Techniques for feedback pronunciation

The techniques for feedback pronunciation can be divided into six types: explicit correction, recast, elicitation, metalinguistic feedback, clarification request, and repetition [11]. The effectiveness of methods is the object of the study. The automatic systems of computer-aided pronunciation training occupy the treacherous ground because of the novelty of the technology and continuously changes of computer systems. The findings of the study shows that there are severe pedagogical deficiencies in many available computer-aided language learning systems, the computer-aided speech recognition systems can be used effectively in learning process.

VI. Conclusion

A core principle of communicative language learning is that the knowledge of syntax and vocabulary form is only a part of a larger hierarchy. Evaluating student's communicative competence is a major research challenge.

By teaching English we often face the situations where a student who knows well grammar fails in managing every day situations in real world. The classroom activities should leads to the progress. Language instructions based on communicative principles are the best way to form student's ability to interact with each other and more comprehensive examinations must be performed to measure student progress [10]. Intelligible pronunciation is only one of the needed skills for speaking a foreign language, and it is often not emphasized in the classroom.

A foreign language learner will make a number of pronunciation errors at the phonemic (segmental) and prosodic levels when producing speech in a target language. Errors at the segmental level can be generally classified as substitution, insertion, deletion, and duration errors. Errors at the prosodic level are more difficult to categorize. There is some debate over whether phonetic or prosodic aspects of pronunciation have more impact on perceived pronunciation quality [12].

Practically, to know a foreign language means to generate the skills and develop the ability to think as a native speaker and to understand other people's thoughts. In order to sounds right and intelligible to the listener the utterance should be conveyed into correct intonational pattern. This means that the internal or external performance of speech should be presented with an appropriate dynamic acoustic connotation in accordance to the rules of the target language.

Computer-aided intonation training is specifically designed to evaluate and improve pronunciation in foreign languages. Due to computer-aided intonation training system the specific pronunciation mistakes will be identified at the word or subword level, providing an opportunity to improve pronunciation in and outside the classroom according to visual feedback.

References

[1] H.Y. Antipova. Guide to English intonation: (in English) / Student's book for pedagogical institutes and faculties in foreign languages./H.Y.Antipova, S.L.Kanevskaya, G.A.Pigulevskaya, - 2d Edition, - M.: Education, 1985. - 224 pp.

[2] Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E. (2006) Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis // Speech Prosody: Proceedings of the 3rd International conference. Dresden, Germany: Vol. 2. – pp. 553-556.

[3] Ockenden M., (2005) Situational Dialogues // The English Centre, Eastbourne / Revised Edition. - Longman - 98 pp.

[4] Gimson A.C. Inoduction to the Pronunciation of Eglish, London, 1966, p. 261.

[5] John H Underwood. Linguistics, Computers and the LanguageTeacher: A communicative Approach. Newbury House Publishers, Inc., Rowley, MA, 1984.

[6] R.S. Hart. The Illinois PLATO foreign languages project. CALICO journal, 12(4):15- -37, 1995.

[7] L. Neumeyer, H. Franco, V Abrash, and L Julia. WebgraderTM: a multilingual pronunciation practice tool. In Proceedings of ESCA Workshop on Speech Technology in Language Learning, 1998.

[8] L. Neumeyer, H. Franco, V Abrash, and L Julia. WebgraderTM: a multilingual pronunciation practice tool. In Proceedings of ESCA Workshop on Speech Technology in Language Learning, 1998.

[9] Marikka Elizabeth Rypa and Patti Price. VILTS: A Tale of Two Technologies. CALICO journal, 16(3):385–404, 1999.

[10] N. Mote, L. Johnson, Abhinav Sethy, Jorge Silva, and S. Narayanan. Tactical language detection and modeling of learner speech errors: The case of Arabic tactical language training for American English speakers. In Proceedings of InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems, page 19, 2004.

[11] R. Lyster and L. Ranta. Corrective feedback and learner uptake. Studies in Second Language Acquisition, 19:37–66, 1997.

[12] Murray J Munro and Tracey M Derwing. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. Language Learning, 49(S1):285–310, 1999.

ЛИНГВИСТИЧЕСКИЕ И АКУСТИЧЕСКИЕ РЕСУРСЫ КОМПЬЮТЕРНОЙ СИСТЕМЫ ОБУЧЕНИЯ ДИАЛОГОВОЙ ИНТОНАЦИИ АНГЛИЙСКОЙ РЕЧИ

Здоронок Ю.А.

В этой статье рассматриваются супрасегментные феномены речи, с которыми связана не только правильная артикуляция звуков изучаемого языка, но и грамотное произношение, просодическая оформленность мысли в речи с помощью эталонных интонационных шаблонов. Лингвистические и акустические ресурсы

необходимы для компьютерной системы обучения диалоговой интонации английской речи в пределах и за пределами учебной аудитории.