**DOI** 10.15622/sp.59.8

M. SEČUJSKI, S. OSTROGONAC, S. SUZIĆ, D. PEKAR

# LEARNING PROSODIC STRESS FROM DATA IN NEURAL NETWORK BASED TEXT-TO-SPEECH SYNTHESIS

*Sečujski M., Ostrogonac S., Suzić S., Pekar D.* **Learning Prosodic Stress from Data in Neural Network based Text-to-Speech Synthesis.**

**Abstract.** Naturalness is one of the most important aspects of synthesized speech, and state-of-the-art parametric speech synthesizers require training on large quantities of annotated speech data to be able to convey prosodic elements such as pitch accent and phrase boundary tone. The most frequently used framework for prosodic annotation of speech in American English is Tones and Break Indices – ToBI, which has also been adapted for use in a number of other languages. This paper presents certain deficiencies of ToBI when applied in synthesis of speech in American English, which are related to the absence of tags specifically intended to mark differences in the level of prosodic stress (emphasis) related to a particular sentence constituent. The research presented in the paper proposes the introduction of a set of tags intended for explicit modeling of the degree of prosodic stress. Namely, a certain sentence constituent can be particularly emphasized, when it is the intended focus of the utterance, or it can be de-emphasized, as is commonly the case with phrases reporting direct speech or with comment clauses. Through several listening tests it has been shown that learning such prosodic events from data has distinct advantages over approaches attempting to exploit the existing ToBI tags to convey the degree of emphasis in synthesized speech. Namely, speech synthesized by a neural network trained on data tagged for the level of prosodic stress appears more natural, and the listeners are more successful in locating the sentence constituent carrying prosodic stress.

**Keywords:** American English, prosodic stress, speech synthesis, ToBI.

**1. Introduction.** The quality of text-to-speech (TTS) synthesis systems is generally rated in terms of the intelligibility and the naturalness of the speech they produce. The intelligibility of synthesized speech is a well-defined concept, which is also easily evaluated through measures such as tests based on semantically unpredictable sentences (SUS) [1]. On the other hand, naturalness is a less defined concept, but it has nevertheless been widely used as a measure of TTS quality at events such as Blizzard challenges [2, 3]. The perceived feeling of naturalness of synthetic speech is based on a number of parameters that are difficult to identify and enumerate, and consequently, listeners are unable to tell what exactly contributes to naturalness [4]. Although there is no general consensus as to what naturalness is, a number of parameters related to it have been proposed, ranging from the ease of comprehension to the internal coherence of the acoustics of the utterance [5]. In many studies

the general quality of speech, or its similarity to natural human speech, is the only concept of naturalness that is evaluated. Ultimately, a successful text-to-speech (TTS) system should be able to convince listeners that they are listening to actual human speech.

Although synthetic speech has reached the level of intelligibility needed for wide practical application a long time ago, there are still challenging problems that remain to be solved. The current focus of the TTS research community is the synthesis of expressive content, which includes emotional expressivity, synthesis of different speaking styles, but no less importantly, synthesis of prosodic elements that convey linguistic meaning. Namely, the prosodic features of a natural-sounding synthesized utterance (the fundamental frequency contour — $f_0$, durations of phonetic segments, as well as temporal changes in volume) should match the features in a possible rendition of the same utterance by an actual human speaker, having in mind that there are many possible renditions of a single utterance, but that some of them may indicate differences in meaning. From the point of view of the listener, the main purposes of prosody in synthetic speech is to indicate syntactic boundaries and reveal some of the underlying syntactic structure of the utterance, as well as to facilitate the recognition of sentence constituents by exploiting the linguistic function of intonation and stress through combining different prosodic variables – pitch, length, loudness and timbre (quality of sound). The variability of these factors in speech appears to be largely ignored by the listeners. However, when some of them are missing or are inadequate, this is perceived as unnatural, and can even impair the intelligibility of synthesized speech, particularly in languages with stress or accent minimal pairs (pro-*test* vs. *pro*-test in English).

Nowadays, users of state-of-the-art dialogue systems expect to interact along the same principles that they use when interacting with other human beings, and consequently, dialogue systems are expected to behave and speak like human beings [6]. There have been various directions of research into how synthesized speech can be made more human-like. For instance, adding non-verbal elements such as laughing, breathing and clicking noises has been shown to increase the user's perception of naturalness of synthetic speech [7]. There has also been significant research effort aimed at investigating the influence of the insertion of filled pauses [8] or other manifestations of hesitation disfluency [9]. However, much of the naturalness of synthetic speech is

ruled by factors with deeper linguistic roots. Namely, for the user of a speech synthesis system to receive information with minimum cognitive effort, it is important that the system should be able not only to provide basic prosodic cues such as word stress or pitch accent to the listener, but to be able to convey elements such as rising intonation that turns a statement into a yes/no question, or prosodic stress, i.e. placing of emphasis on particular words because of their relative importance in the sentence.

Prosodic stress is often used pragmatically to focus the attention of the listener on particular words or the ideas associated with them, thus changing or clarifying the meaning of a sentence:

- *John* met Iris today. (Iris wasn't met by someone else today.)
- John met *Iris* today. (John didn't meet someone else today.)
- John met Iris *today*. (John and Iris didn't meet on some other day.)

Prosodic stress typically manifests itself as an increase in the prominence of stressed syllables, in terms of one or more prosodic variables previously mentioned. Words associated with prosodic stress are usually pronounced with louder and longer stressed syllables, and their fundamental frequency (pitch) usually extends over a wider range [3]. Stressed vowels in words carrying prosodic stress are typically associated with a more prominent pitch or pitch movement, increased duration and loudness, and they tend to be more peripheral in quality then vowels which are not associated with prosodic stress, which are normally more centralized. Furthermore, the stress-related acoustic differences between the syllables of a word that is not prosodically stressed are generally small compared to the differences between the syllables of a word which carries prosodic stress (cf. e.g. *Iris* and *today* in the examples above). Another use of prosodic stress is related to stress patterns that can be typical of a certain language, e.g. in French prosodic stress is typically placed on the final syllable of a string of words. This research will principally deal with the pragmatic use of prosodic stress, which is also referred to as contrastive stress. In natural human speech, there are also words and entire phrases that are pronounced in a pitch range that is compressed, in order to indicate that they are less relevant or that they do not bring any new information to the listener. A speech dialogue system that aims at establishing effortless speech communication with a human user should be able to provide such linguistic cues as well.

Parametric speech synthesizers represent the most widely used speech synthesis technique today, owing to their capability to learn complex mappings from linguistic features to acoustic features from data. They usually require large quantities of speech data to learn from, and obtain best results if the speech data is properly annotated. The principal task of a parametric speech synthesis system is to convert the input text into the acoustic features of speech which will be produced by a vocoder. To make this task easier, the input text is usually accompanied by annotation at various levels, not only regarding the phonemic identity of phonetic segments, but also prosodic features, at the level of syllable, word, phrase, or the entire utterance. On the other hand, the annotation of speech corpora is known to be an extremely time consuming task, requiring a lot of human effort, and often requiring the engagement of expert linguists.

Phonetic annotation represents the marking of phoneme and word boundaries and it can be carried out automatically with relatively high accuracy, based on the alignment of phonetic transcriptions and speech data collected from the voice talent. In order to avoid possible training errors introduced by faulty phonetic transcription, manual verification of phone and word boundaries is usually performed, possibly aided by a suitable graphical user interface. On the other hand, prosodic annotation represents the marking of a range of prosodic events at different levels, and is most often entirely manual. Prosodic annotation is carried out according to a chosen intonational model, which attempts to describe the intonation and temporal structure of the sentence. Intonational models can be divided into two broad categories, depending on the way they treat the dynamic character of the speech signal [1]. *Phonetic* models of intonation attempt to provide an explanation of the intonational features of the speech signal, especially the fundamental frequency. However, they are principally based on physical features and as such are unable to provide a connection to a discrete set of linguistic features that have a great influence on the acoustics of the utterance. *Phonological* models, on the other hand, are directly relevant to the listeners and their perception of speech, as they establish the relationship between the acoustic features of the signal and a corresponding discrete set of linguistically motivated prosodic events. For these reasons phonological models are relevant to both automatic speech synthesis and recognition. The intonational model most often used for American English is the Tone and Break Indices (ToBI), which is based on indexing pitch accents, phrase accents as well as boundary tones [11].

The remainder of the paper is organized as follows. The next section gives a brief overview of the standard ToBI model for American English, followed by a discussion on some of its shortcomings related to the synthesis of expressive speech. The issue of prosodic stress as well as reproduction of utterances containing direct speech and reporting phrases are given particular attention. To overcome these shortcomings of ToBI, the study described in this paper proposes an extension to the standard set of ToBI tags, which consists of the introduction of explicit marking of the degree of emphasis that the speaker associates with particular sentence constituents. Section 3 will present an experiment involving a listening test performed by 20 listeners, which confirms that the use of the ToBI model extended in this way leads to an improvement in the naturalness of synthesized speech and allows the listener to estimate the relative importance of particular words or phrases more accurately. The following section discusses the results of the experiment, while the concluding section summarizes the paper and provides an overview of the directions of future research.

**2. ToBI intonational model and its shortcomings**. Tone and Break Indices (ToBI) is a high-level prosodic model that has firstly been developed for American English, and was later extended with a number of variants for other languages [12]. ToBI represents the intonation of an utterance as a linear concatenation of tonal events, and global intonational contours are explained as concatenations of local strings of events.

A ToBI prosodic transcription of a particular utterance describes its tonal events and internal phrase structure, and can also provide other information as well. The term *tonal event* includes pitch accents, phrase accents as well as boundary tones. Tonal events represent combinations of high and low tones that may be associated with stressed syllables. The pitch accent that will be used in a certain situation depends largely on the syntax of the utterance, but it can also depend on semantics as well as a specific intention of the speaker. Consequently, the confidence with which pitch accents can be predicted is much lower than e.g. the confidence with which one can predict stressed syllables within a word. Since within each pitch accent a stressed syllable can be assigned a high or a low tone, pitch accents are divided into two groups – high (such as H*) and low (such as L*), and various other combinations such as L+H* i L*+H, are also allowed, with asterisk indicating the stressed syllable. Since the speaker assigns pitch accents i.e. prosodic prominence only to words which he/she considers important in a given situation, it is also possible that a stressed syllable does not carry a pitch accent at all.

The ToBI model also uses appropriate tags to indicate the internal phrase structure of an utterance, although the model is still essentially linear. Phrase breaks are indicated with levels from 0 to 4, where e.g. the lowest-level break index (0) is defined in terms of connected speech processes (occurring at boundaries such as "*did you*"), 1 indicates the typical absence of break at most phrase-medial word boundaries, while 4 indicates the boundary between two full intonational phrases. Namely, the utterance is divided into intonational phrases, delineated by level 4 breaks, and within an intonational phrase it is possible to identify intermediate phrases, delineated by level 3 breaks. Break index 2 indicates a sense of disjuncture at a boundary between two words where there is no acoustic evidence of tonal events and thus break indices 3 and 4 are not suitable. Each intermediate phrase boundary (break index 3) is assigned a phrase accent (L-, !H- and H-, where the exclamation mark denotes a downstep related to the previous H), while each intonation phrase (break index 4) is assigned a boundary tone (L% and H%). As each intonational phrase consists of at least one intermediate phrase, each level 4 boundary represents at the same time a level 3 boundary, which means that there are 6 possible combinations of phrase accents and boundary tones that can appear at the end of an intonational phrase (L-L%, H-L%, !H-L%, L-H%, H-H% and !H-H%). Phrase accents are also used to indicate the beginnings of intermediate phrases (%H and %L). The standard ToBI system includes other tags, such as the diacritic '<', which is used in combination with a high pitch accent when the syllable that follows the stressed syllable is higher than the stressed syllable (delayed peak). Some of the tags used by standard ToBI have been ignored in this research for being irrelevant (e.g. the tags indicating disfluencies in spontaneous speech were not used since the speech corpora used for training contain only fluent speech), and optional ToBI tags were not used.

Local tonal events, i.e. pitch accents, phrase accents and boundary tones, are considered as targets along the global intonational contour, and standard ToBI assigns local tonal events to specific points in time. However, this relationship between ToBI tonal events and particular time instants should be considered rather loose, since any identification of temporal events in linguistically motivated abstract representations would essentially be meaningless. In order to emphasize the symbolic aspect of ToBI, as well as to further simplify both ToBI corpus annotation and prediction of ToBI tags in synthesis, in this research it was assumed that

ToBI tags are assigned to particular phones, or in some cases words, instead of being assigned to arbitrary points in time. Similar modifications to the standard ToBI framework have already been implemented in speech synthesis systems such as *Festival* [13].

The standard ToBI model possesses many advantages which make it the model of choice for use within speech technology systems, and it has been defined having in mind its possible use within such systems. Namely, prosodic events are clearly defined, in a way that can be easily interpreted by a computer, the model is easily extensible to other languages and linguistic phenomena, and it has been designed so as to minimize the degree of disagreement between different ToBI annotators [14]. A speech synthesis model such as a neural network, when trained on ToBI annotated speech, is able to learn to reproduce the acoustic features of speech from its ToBI annotation. However, it should be kept in mind that the conversion of an intonation contour into its ToBI representation is many-to-one, which means that a number of intonation contours, which can be vastly different among themselves in terms of absolute frequencies and temporal behaviour, can correspond to the same ToBI transcription. Consequently, even within a single speaker, a ToBI transcription cannot be uniquely mapped into a set of acoustic features, i.e. a parametric speech synthesizer could construct a number of different sets of acoustic features based on a single ToBI transcription. As regards speech synthesis, it is sufficient that a parametric speech synthesizer should produce a set of acoustic features which would yield speech that sounds acceptable in a particular context.

From the point of view of expressive speech synthesis, an important advantage of using ToBI for symbolic representation of prosody is that it is possible to control some of the prosodic features of synthesized speech by manipulating its ToBI transcription. For instance, by manually changing pitch accents, phrase accents and boundary tones, statements can be turned into yes/no questions, and it is also possible to direct the attention of the listener to a particular word in the utterance. On the other hand, the ToBI model also has a number of disadvantages. Firstly, even for a person with relevant linguistic knowledge, ToBI annotation is a rather complex process, which has been reported to last up to several hundred times more than the duration of the speech being annotated [15]. Furthermore, since ToBI annotation essentially relies on the annotator's tacit understanding of the relationship between the objective intonational contour and its symbolic representation, there

is a strong subjective component to it. Consequently, regardless of the original intention to minimize the inter-annotator disagreement, it is still reported to be relatively high [16]. Finally, due to the complexity of the ToBI tagset, some tags or tag combinations may be scarce or completely absent from speech corpora, which has a negative effect on training [14]. For that reason, common modifications of the standard ToBI model usually include merging some of its categories into one, as it was done e.g. in [17].

The weakness of the standard ToBI model which is of particular relevance for this research is that it is not quite suitable for conveying linguistically relevant prosodic features such as prosodic stress. Namely, the relationship between ToBI tags and the perception of importance by the listener is defined on a relative scale, which means that ToBI can convey that one word may be more prosodically prominent than another, but cannot convey the absolute degree of its prosodic prominence. The distinction that ToBI makes between words with pitch accents and words without pitch accents appears to be insufficient to indicate e.g. that, among the words with pitch accents, one is significantly more important than others.

As far as the prosodic stress (assigning particular prominence to a word which carries new or particularly important information) is concerned, as in:

*Sarah will go to **London** in September.* (1)

ToBI does not offer an explicit solution. Clearly, the novelty of the information conveyed is related to the ToBI tags used, and this relationship has been the topic of much research. For example, in [18] the use of H* and L+H* pitch accents is reported to be related to novelty, while given or available information is assigned other pitch accents, depending on the context. However, this relationship is not conclusive enough to serve as a basis for the reproduction of prosodic stress in speech synthesis. For example, an instance of prosodic stress indicated with a bitonal accent L+H* in the synthesis of expressive speech is not necessarily converted to an acoustic representation characterized by sufficient prominence of the stressed syllable so as to unambiguously indicate prosodic stress. The most common reasons for this are related to the lack of training data, as well as the fact that the acoustic realizations of prosodic stress may be highly variable, and

commonly affect not only the intonation contour, but also the duration of particular phonetic segments as well as the manner of their articulation. For all these reasons, in practice more reliable means for conveying emphasis are usually preferred. For instance, to indicate the prosodic prominence of a particular word, the IBM speech synthesizer [19] combines the pitch accent H* and the phrase accent L-. The motivation for introducing suck a one-to-one mapping between contrastive prosodic stress and particular ToBI tags came from the analysis of a very small corpus consisting of several declarative sentences spoken by 20 professional speakers. It was found that the contrastively-emphasized word consistently had at least intermediate prosodic phrase boundaries on each side of the word, accompanied by break indices of level at least 3. However, although such a representation may have provided an unambiguous cue to prosodic stress, a one-to-one mapping between prosodic stress and particular ToBI tags is not what happens in practice, and thus results in a certain loss of naturalness. The research described in this paper exploits the appearance of more powerful automatic learning algorithms which are able to establish a more sophisticated relationship between prosodic stress and its acoustic counterparts. For that to be possible, an explicit tag (E+) has to be introduced in order to indicate that the speaker has intentionally emphasized a particular word. By training on a speech corpus that contains words tagged with E+ it was made possible for the system to establish, by learning from data, the connection between the intention to emphasize a word and its acoustic realization.

Similarly, standard ToBI does not offer any possibility to explicitly mark content that is commonly de-emphasized, such as comment clauses or inquit formulas used to report direct speech:

*It would be nice,* ***I suppose****, if they keep their promise.* (2)

*"You should have seen it coming,"* ***I replied****.* (3)

Such phrases are commonly pronounced in a compressed range of fundamental frequency, i.e. pitch, in order to convey their lower degree of importance or the fact that they are not a part of the main clause. The research presented in this paper also investigated the possibility of annotating deemphasized content with a specific tag (CF0 — compressed $f_0$), in order to allow the system to reach its own conclusions as to the relationship between

the intention to de-emphasize a clause and its acoustic realization. To the best of our knowledge, there has been no research effort to explicitly model this type of dependency or introduce reduced emphasis into synthesized speech. An example of a ToBI annotation following the guidelines modified so as to accommodate for the introduction of E+ and CF0 tags is given in Figure 1.

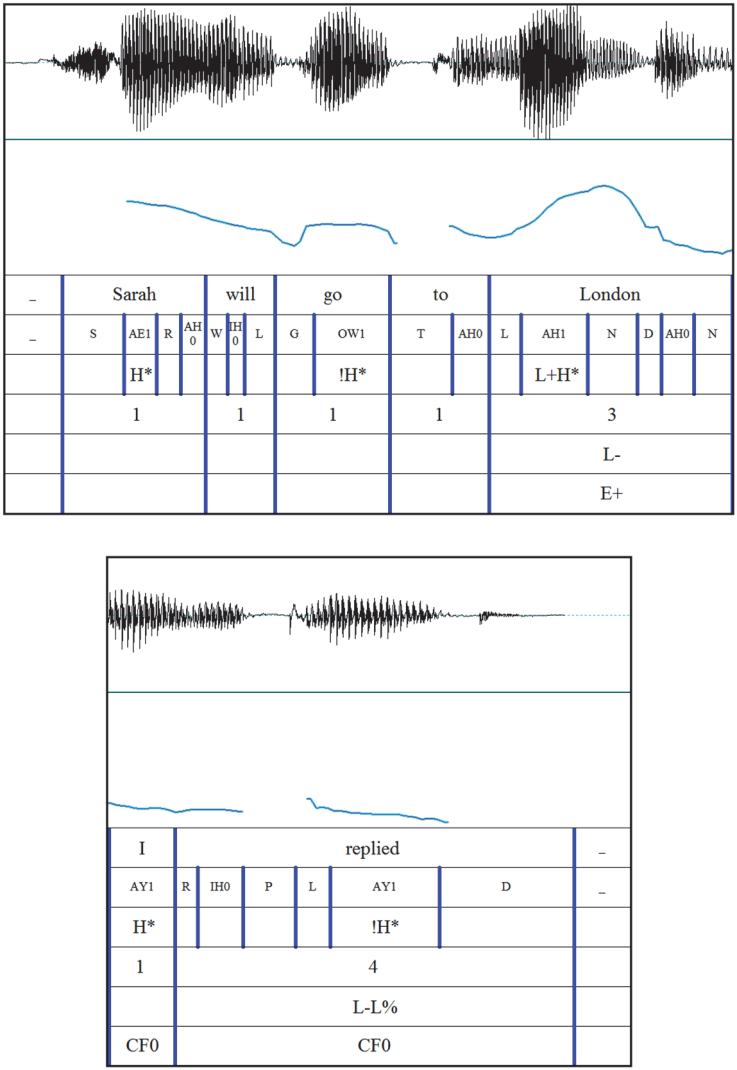


Fig. 1. Prosodic annotation of the sentence "*Sarah will go to London, I replied.*"

The same guidelines were used to annotate all speech data used for training in this research. Namely the annotators were instructed to do the following:

– Use the E+ tag on words whose prosodic features, including not only intonation but also the degree of articulation effort, seem to indicate that the speaker intended to assign particular importance to the word because it brings new information or is contrasted to an alternative word (stated or implied). This annotation was used on top of ToBI with no restrictions, in order to make it ToBI independent. For instance, if the articulation effort was strong enough, even the words regularly tagged with !H* could be assigned E+. Furthermore, while E+ was commonly indicated by higher pitch values in words with high ToBI pitch accents, it was also indicated by lower pitch values in words with low pitch accents. Such a use of E+ allowed it to be effectively excluded from some rounds of experiments, as it will be explained in the following section.

– Use the CF0 tag on clauses that are pronounced in a compressed pitch range, which seems to indicate that the speaker intended to assign to them a lower degree of emphasis than to the remainder of the utterance. These cases most notably included, but were not limited to, phrases reporting direct speech, comment phrases, asides and right dislocations related to afterthoughts. There is, however, a certain degree of dependence of ToBI annotation on the presence or absence of CF0. Namely, the clause under CF0 was ToBI annotated as if its pitch range was normal, which means that simple removal of CF0 from the speech corpus may not be adequate in all cases, unlike the case with E+.

**3. The experiments**. The experiments have been carried out on two speech corpora of American English, containing utterances provided by one male speaker (M) and one female speaker (F), both professional voice talents. The basic data related to the speech corpora is given in Table 1, including the data on the total number of E+ and CF0 tags. Both E+ and CF0 tags were used at word level, so it should be noted that the number of CF0 tags indicates the total number of *words* in clauses tagged with CF0. The numbers of E+ and CF0 tags are not the same across the two speakers, but this difference was disregarded in the experiment and two synthesizers, each trained on one of the corpora, were used in the listening tests in equal measure, i.e. the listeners were provided with examples of synthesized speech from both, male alternating with female for diversity. The results of the experiments were analyzed without regard

to the fact which of the corpora was used for training the system that provided a particular example of synthetic speech.

Table 1. Content of speech corpora

| | M | F |
|---|---|---|
| Duration | 3h 33min | 4h 20min |
| # of utterances | 3316 | 4556 |
| # of words | 43632 | 49580 |
| # of E+ | 629 | 1162 |
| # of CF0 | 2359 | 5522 |

The model used for synthesis of utterances in all experiments was based on deep neural networks (DNN), owing to the fact that they clearly outperform previous parametric approaches to speech synthesis [20]. The model is described in detail in [21], and it will be briefly presented here. It was developed using the *Merlin* toolkit [22] with some modifications, as well as the CNTK framework [23]. The WORLD vocoder [24] is used to convert the acoustic features provided by the model into a speech signal, and is also used to provide the acoustic feature vectors in the training phase.

In the synthesis phase, the input text is firstly processed to obtain linguistic features relevant for synthesis. The obtained linguistic features include phone identity and the identities of neighbouring phones, the phone position related to syllable/word/phrase boundary, word position related to phrase boundary, number of phones in syllable/word, number of words in phrase/utterance, part-of-speech information as well as prosodic information represented by ToBI transcription. After the linguistic features are obtained, the acoustic features are produced from the phonetic transcription of the text augmented with the obtained linguistic features. The segment of the speech synthesis system charged with the production of acoustic features from the phonetic transcription and linguistic features of input text consists of two deep neural networks — the duration network and the acoustic network, as shown in Figure 2.

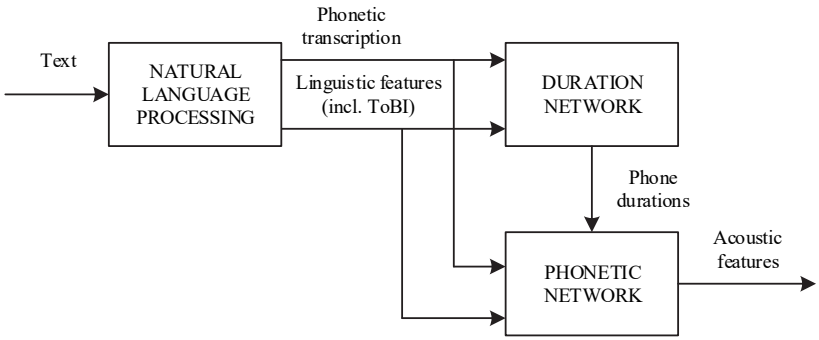


Fig 2. Neural network based model used for speech synthesis

The first network models phoneme durations and the second one models context-dependent acoustic features. The input for both networks is the phonetic transcription of the input text accompanied by linguistic features. In the training phase, the duration network adjusts its weight coefficients by minimizing the objective distance between the predicted values of HMM state durations of a phoneme and their actual durations in the training speech corpus. The actual values of HMM state durations of phonemes, as target features for the duration network, are extracted by forced alignment from training data, following the procedure proposed in the *Merlin* toolkit [22]. The inputs and outputs of the duration network are phone aligned. At synthesis phase, the duration network is required to predict the HMM state durations, and hence the durations of phones themselves, based on phone identity and linguistic context defined by features mentioned above. The HMM state durations obtained at the output of the duration network at synthesis time are used as additional inputs to the acoustic network. In the training phase, the acoustic network is trained to predict the relevant acoustic features given the phonetic transcriptions, linguistic context and HMM state durations. During the training, the acoustic network uses acoustic features extracted from speech recordings by the WORLD vocoder as target features. The acoustic features used include mel-generalised cepstral coefficients (MGCs), band aperiodicities (BAP) as well as log $f_0$, and they are further extended with their first and second derivatives as well as an additional flag indicating whether the current frame is voiced or unvoiced (V/UV). In all experiments in this paper 40 MGCs, 1 BAP, 1

log $f_0$ and 1 V/UV feature are used, yielding output feature vectors of length 127. Since the inputs and outputs to the acoustic network are frame aligned, the input feature vector of the acoustic network, in both training and synthesis phases, is extended by additional numeric features, including the index of the current frame in the state/phone as well as the index of the current state. Both networks consist of 4 hidden layers with 1024 neurons each. The first three have tangent hyperbolic as the activation function, while the fourth layer is recursive and uses long short-term memory (LSTM) neurons. The output layer is linear. The objective function used is mean squared distance between the predicted and the actual values in the training data. The input features are normalized to the interval [0.01, 0.99], while the output features are *z*-normalized. Each of the two networks is separately trained by backpropagation and stochastic gradient descent optimization. The smoothness of static features is achieved by using the maximum likelihood parameter generation algorithm [25], taking into account the predicted dynamic features. After the formants are further enhanced by postfiltering, the acoustic features are fed to the WORLD vocoder in order to generate speech waveforms.

As previously mentioned, two versions of the synthesizer were built, each trained on one of the two available speech corpora. Therefore, it was possible to synthesize sentences in either of the voices, M or F, based on specified phonetic transcription and ToBI annotation. This approach allowed the control over certain aspects of sentence intonation related to emphasis, as will be explained in more detail below.

The listening tests included 20 listeners who were not native speakers of English, but who professed to possessing good English language skills. The listeners had no or little previous experience with testing speech technology systems. The tests have been carried out in a relatively silent environment, using high-quality headphones. Nevertheless, it should be noted that the quality of the reproduction of synthesized speech is not of primary importance here because all experiments focus on prosodic features that are relatively robust to impairments in signal quality. Each of the listeners was required to evaluate 22 sentences, and each of the sentences was synthesized in 3 versions:

– Version A: Both ToBI tags E+ and CF0 were used both in training and in synthesis, as proposed by the research and described in the previous section.

– Versions B and C: Neither E+ nor CF0 were used in either training or synthesis, which corresponds to the standard ToBI model. In both cases, the effect of E+ was simulated by using either H* or L+H* in the ToBI transcription that is used as DNN input, as suggested by the findings of [18]. In the version B the effect of CF0 is ignored (i.e. the clause in a compressed $f_0$ range is annotated in the regular way, as if it was in no way different from the remainder of the utterance), while in the version C the pitch accents are removed from all words which were annotated with CF0 in version A. In this way the CF0 tag was equalized with the absence of a pitch accent. All of the aforementioned methods of simulating E+ or CF0 are in accordance with standard approaches based on the ToBI model.

The differences between versions A, B and C are conveniently summarized in Table 2.

Table 2. Versions of synthesis used in the experiments

<table>
<tr><th>A</th><th>B</th><th>C</th></tr>
<tr><td rowspan="2">E+ and CF0 used both in training and synthesis</td><td colspan="2">E+ and CF0 not used in either training or synthesis, E+ simulated using H* or L+H*</td></tr>
<tr><td>CF0 ignored</td><td>Pitch accents removed from words that have CF0 in version A</td></tr>
</table>

*Experiment 1.*

(a) In 10 sentences similar to the one from Example 1, with one word carrying the semantic focus of the sentence, the task of the listener was to determine, given 4 options, which word was assigned prosodic stress (E+) by the synthesizer, i.e. which of the words was intended to be emphasized. In none of the cases was it possible to determine the emphasized word solely on the basis of textual content of the utterance. The utterances were presented visually to the listeners, with available options indicated in boldface, as in:

***Sarah talked*** *to her* ***neighbour*** *about a* ***problem***.

(b) In the second part of the experiment the same 10 sentences were used, but now the listeners were told which of the words was the intended focus of the utterance, and they were required to grade (on the scale from 1 to 5) how successfully this was conveyed in synthesized speech.

The aim of the Experiment 1 was to establish the effect of introducing E+ as opposed to signalling emphasis by rule-based methods. For that reason, the 10 utterances offered to the listeners included 5 utterances synthesized according to version A and 5 utterances synthesized according to version B. To minimize the influence of factors over which we had no control and which may have influenced the results of the listening tests, both the word which was assigned E+ and the order of presentation of the utterances were randomly varied across the listeners. For instance, if a word to which an E+ tag was assigned also happened to carry an L- phrase accent, the impression of emphasis was increased by the phrase accent, which would obscure the actual influence of E+. To mitigate this effect, versions where E+ was assigned to all 4 candidate words were used in the experiment in equal measure.

*Experiment 2.*

(a) In 12 utterances similar to those from Examples (2) and (3), including reporting clauses or comment clauses, which are commonly delivered in a compressed $f_0$ range, the listeners were required to grade (on the scale from 1 to 5) the general naturalness of intonation in synthesized speech.

(b) Each of the 12 utterances was presented to the listeners in all 3 versions, and they were required to select the version with most natural intonation.

The aim of the Experiment 2 was to establish the effect of introducing the CF0 tag on the naturalness of intonation, as opposed to signalling a lower degree of emphasis by excluding pitch accents or not signalling it at all. For that reason, both parts of the experiment included examples from all 3 synthesis versions in equal measure, in a random order unknown to the listeners. The listeners were unaware of the aim of the Experiment 2, but it was nevertheless possible for them to conclude that this experiment is related to a variable degree of emphasis associated with reporting or comment clauses.

**4. Results and discussion**. The results of all experiments are shown in Figure 3. The results of Experiment 1, concerned with E+, show that the listeners have been consistently identifying the sentence focus

most successfully in the version A of synthesis (76.0%, as opposed to 41.0% for version B, while 25.0% would correspond to a random choice in both cases). They were also consistent in assigning version A higher marks when judging how successfully the sentence focus was conveyed in synthesized speech (4.17 on average, as opposed to 3.35 for version B). As regards the CF0 tag, results are less conclusive because in some cases the difference between specific synthesis versions was almost negligible, as reported by a majority of listeners.

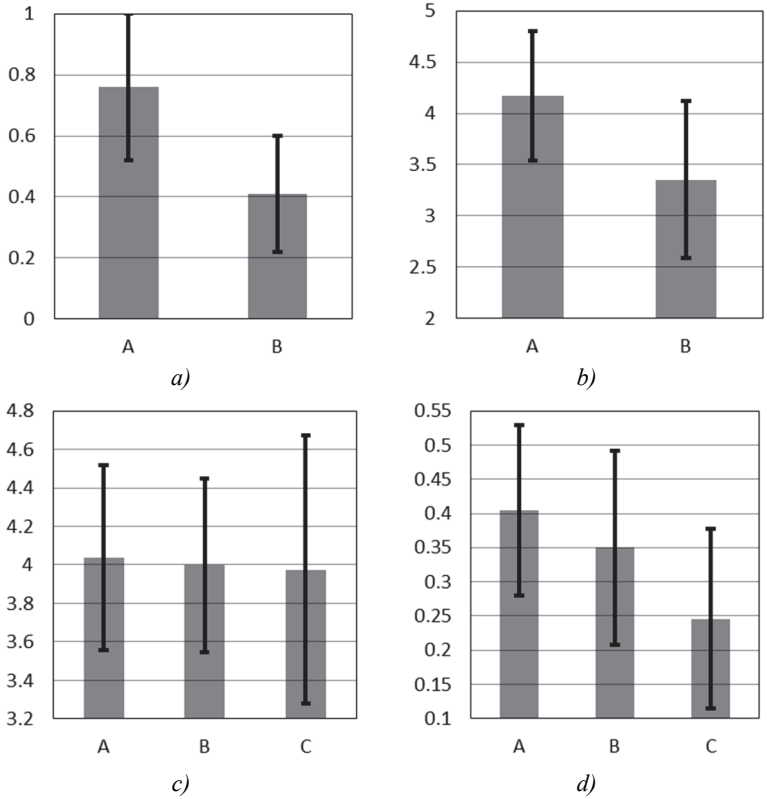


Fig. 3. Experiment results (mean value and standard deviation): a) Experiment 1a: percentage of correct identification of the emphasized word; b) Experiment 1b: average grade; c) Experiment 2a: average grade; d) Experiment 2b: relative frequency of a particular version being chosen as the most natural one

In the Experiment 2a, when evaluating utterances individually, the tendency towards assigning higher marks to the version A was practically

negligible (4.04, as opposed to 4.00 and 3.97 for versions B and C respectively). However, in direct comparison in Experiment 2b, the version A was noticeably more often identified as the most natural one (in 40.4% cases, as opposed to 35.0% and 24.6% for versions B and C respectively, while in this case 33.3% would correspond to a random choice). Furthermore, it can be noted that version C was least often recognized as the best one in direct comparison (Exp. 2b), although in individual evaluation (Exp. 2a) it received approximately the same marks as the version B.

Based on the results, it can be concluded that the initial hypothesis has been confirmed by the experiments, and it can also be noted that the variance between the answers given by particular listeners was significantly higher in case of CF0. A possible explanation of this fact is that the listeners may have had different expectations with respect to the delivery of reporting phrases, comment phrases or other content commonly delivered in a compressed $f_0$ range. For example, even when a reporting phrase is not synthesized within a compressed $f_0$ range, it can still sound quite acceptable to the listener, especially in a listening test, as opposed to actual speech communication. On the other hand, the expectations of listeners as regards prosodic stress are relatively clear and unambiguous. This may explain a considerably greater variance between the grades given by different listeners in Experiment 2 than in the Experiment 1. It should also be noted that the reproduction of E+ and particularly CF0 by the neural network based synthesizer in some cases was not entirely adequate. It can, thus, be concluded that of the reasons why the experiment results did not quite meet the expectations is certainly the inability of the DNN model to faithfully reproduce the E+ and CF0 tags after being trained on the available quantity of speech data.

**5. Conclusion**. The paper has presented a research aimed at increasing the quality of synthesis of expressive speech based on more adequate modeling of linguistically relevant prosodic features of speech, including prosodic stress and delivery of speech in a compressed $f_0$ range. In all cases of interest it has been shown that, if the standard ToBI model is augmented by tags aimed at reproduction of prosodic stress (E+) and content delivered in a compressed $f_0$ range (CF0), the target prosodic feature is more easily identified and there is an increase in overall naturalness.

Although the experiment was primarily concerned with synthesis of American English speech, the universality of the ToBI model suggests that

it would be possible to obtain the same results for other languages for which a ToBI model has been developed, and even for languages for which such a model could be developed in the future. The directions of our future research include the verification of the same hypotheses for Serbo-Croat, another language for which a ToBI model has been developed [26], as well as an investigation into the same phenomena in non-neutral speech styles. Since the range of fundamental frequency constitutes one of the most important features of a speech style or emotional state, it is an important research question to which extent the obtained results apply in case of different speech styles or emotional states.

It should be noted, however, that direct comparison of the results between different studies of this type is difficult because of their language dependence, dependence on corpus size as well as the differences in any of a number of varying parameters, starting from the choice of the system architecture or intonation model. Essentially, the question is to what extent linguistic phenomena should be explicitly modeled. Given a sufficiently complex system architecture and enough data, machine learning systems are able to learn surprisingly complex abstract linguistic concepts. However, as this study suggests, in case of relatively simple systems and a realistic amount of available data, explicit modeling of linguistic factors is still necessary to improve system performance.

## References


1. Dall R., Yamagishi J., King S. Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. Proceedings of Speech Prosody. 2014. 5 p.
2. King S., Karaiskos V. The Blizzard Challenge 2016. Blizzard Challenge Workshop. 2016. 17 p.
3. King S., Wihlborg L., Guo W. The Blizzard Challenge 2017. Blizzard Challenge Workshop. 2017. 17 p.
4. Tatham M., Morton K. Developments in Speech Synthesis. John Wiley & Sons. 2005. 280 p.
5. Sluijter A. et al. Evaluation of speech synthesis systems for Dutch in telecommunication applications. Proceedings of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis. 1998. 6 p.
6. Berg M. Modelling of Natural Dialogues in the Context of Speech-based Information and Control Systems. PhD Thesis. University of Kiel. 2014. 250 p.
7. Trouvain J. Laughing, Breathing, Clicking - The Prosody of Nonverbal Vocalisations. Proceedings of Speech Prosody. 2014. pp. 598–602.
8. Dall R. et al. Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis. Proceedings of the Annual Conference of the ISCA. 2014. 5 p.
9. Székely É., Mendelson J., Gustafson J. Synthesising Uncertainty: The Interplay of Vocal Effort and Hesitation Disfluencies. 18th Annual Conference of the International

Speech Communication Association (INTERSPEECH 2017). 2017. vol. 2017. pp. 804–808.
10. Beckman M.E. Stress and Non-Stress Accent. Foris Publications. 1986. 241 p.
11. Silverman K. et al. ToBI: A standard for labeling English prosody. Proceedings of the 2nd International Conference on Spoken Language Processing. 1992. 4 p.
12. Beckman M.E., Hirschberg J., Shattuck-Hufnagel S. The original ToBI system and the evolution of the ToBI framework. Prosodic typology: The phonology of intonation and phrasing. 2006. 37 p.
13. Black A.W., Hunt A.J. Generating F0 contours from ToBI labels using linear regression. Proceedings of ICSLP. 1996. 4 p.
14. Wightman C.W. ToBI or not ToBI. Proceedings of the International Conference on Speech Prosody 2002. 2002. 5 p.
15. Syrdal A., Hirschberg J., McGory J., Beckman M. Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody. *Speech communication*. 2001. vol. 33. no. 1-2. pp. 135–151.
16. Syrdal A., McGorg J. Inter-Transcriber Reliability of ToBI Prosodic Labeling. Proceedings of the International Conference on Spoken Language Processing (ICSLP). 2000. 4 p.
17. Niemann H. et al. Prosodic processing and its use in Verbmobil. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97). 1997. vol. 1. pp. 75–78.
18. Pierrehumbert J., Hirschberg J.B. The meaning of intonational contours in the interpretation of discourse. Intentions in communication. 1990. pp. 271–311.
19. Hamza W. et al. The IBM Expressive Speech Synthesis System, Proceedings of the Eighth International Conference on Spoken Language Processing (ISCLP). 2004. 4 p.
20. Ze H., Senior A., Schuster M. Statistical parametric speech synthesis using deep neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 7962–7966.
21. Delić T., Sečujski M., Suzić S. A review of Serbian parametric speech synthesis based on deep neural networks. *Telfor Journal*. 2017. vol. 9. no. 1. pp. 32–37.
22. Wu Z., Watts O., King S. Merlin: An Open Source Neural Network Speech Synthesis System. Proceedings of the 9th ISCA Speech Synthesis Workshop. 2016. 6 p.
23. Seide F., Agarwal A. Cntk: Microsoft's open-source deep-learning toolkit. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 2135–2135.
24. Morise M., Yokomori F., Ozawa K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*. 2016. vol. 99. no. 7. pp. 1877–1884.
25. Tokuda K. et al. Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00). 2000. vol. 3. pp. 1315–1318.
26. Gođevac S. Transcribing Serbo-Croatian Intonation. Prosodic Typology: The Phonology of Intonation and Phrasing. 2005. 26 p.

**Sečujski Milan** — Ph.D., associate professor, head of Laboratory of Acoustics and Speech Technology of Faculty of Technical Sciences, University of Novi Sad. Research interests: digital signal processing, speech synthesis, natural language processing, dialogue systems, prosodic modelling, development of speech and language resources, machine learning, neural

networks. The number of publications — 160. secujski@uns.ac.rs; 6, Trg Dositeja Obradovića, 21000, Novi Sad, Serbia; office phone: +381-21-485-2533.

**Ostrogonac Stevan** — senior researcher, AlfaNum – Speech Technologies, software developer, AlfaNum – Speech Technologies. Research interests: text-to-speech synthesis, automatic speech recognition, natural language processing, dialogue systems, development of speech and language resources, machine learning, neural networks. The number of publications — 18. ostrogonac.stevan@alfanum.co.rs; 40, Bulevar Vojvode Stepe, 21000, Novi Sad, Serbia; office phone: +381-64-845-5302.

**Suzić Siniša** — researcher of Laboratory of Acoustics and Speech Technology of Faculty of Technical Sciences, University of Novi Sad. Research interests: expressive speech synthesis, digital signal processing, dialogue systems, machine learning, deep neural networks. The number of publications — 19. sinisa.suzic@uns.ac.rs; 6, Trg Dositeja Obradovića, 21000, Novi Sad, Serbia; office phone: +381-21-485-2521.

**Pekar Darko** — research assistant of the Department for Power, Electronic and Telecommunications Engineering of the Faculty of Technical Sciences, University of Novi Sad, CEO (Chief Executive Officer), AlfaNum Speech Technologies. Research interests: human-computer interaction, speech recognition and synthesis, speaker identification, emotion recognition, speech morphing, numerical simulations, artificial intelligence. The number of publications — 100. darko.pekar@alfanum.co.rs; 40, Bulevar Vojvode Stepe, 21000, Novi Sad, Serbia; office phone: +381-21-485-2521.

**Acknowledgements.** The research is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grants TR32035 and OI178027). The authors are also grateful to the company Speech Morphing Inc. from Campbell, California, USA, for the permission to use their speech databases for this research.

**УДК** 004.5 **DOI** 10.15622/sp.59.8

М. Сечуйски, С. Острогонац, С. Сузич, Д. Пекар

# ОБУЧЕНИЕ ПРОСОДИЧЕСКОЙ МОДЕЛИ ПО ДАННЫМ В НЕЙРОСЕТЕВОМ СИНТЕЗЕ РЕЧИ

*Сечуйски М., Острогонац С., Сузич С., Пекар Д.* **Обучение просодической модели по данным в нейросетевом синтезе речи.**

**Аннотация.** Естественность — один из важнейших аспектов синтезированной речи. Современные параметрические синтезаторы речи требуют обучения на большом количестве аннотированных речевых данных, чтобы иметь возможность передавать просодические элементы, такие как тоническое ударение и фразовый граничный тон. Наиболее часто используемый инструментарий для просодической аннотации речи в американском английском языке — Индексы Тонов и Просодических швов — ToBI, которые также были адаптированы для использования на других языках. В настоящей статье представлены некоторые недостатки ToBI в синтезе речи на американском английском языке, которые связаны с отсутствием тегов, специально предназначенных для обозначения различий в уровне просодии (акцента), связанной с конкретной частью предложения. В данном исследовании предлагается введение набора тегов, предназначенных для точного моделирования степени просодии, а именно определенная составляющая предложения может быть особо подчеркнута, если она является намеченным фокусом высказывания или ее роль преуменьшена, как это обычно бывает с фразами, сообщающими о прямой речи или комментариями.

С помощью нескольких аудирований было продемонстрировано, что изучение просодической модели на основе данных имеет определенные преимущества перед подходами, пытающимися использовать существующие теги ToBI для передачи степени акцента в синтезированной речи: речь, синтезированная нейронной сетью, обученной на данных с тегами уровня просодии, представляется более естественной, и слушатели могут с большим успехом отыскать просодическую составляющую предложения.

**Ключевые слова:** американский английский, просодическая модели, синтез речи, ToBI.

**Сечуйски Милан** — к-т техн. наук, доцент, заведующий лабораторией акустики и речи факультета технических наук, Нови-Садский университет. Область научных интересов: обработка цифровых сигналов, синтез речи, обработка естественного языка, диалоговая система, моделирование интонаций, разработка речевых и языковых ресурсов, машинное обучение, нейронные сети. Число научных публикаций — 160. secujski@uns.ac.rs; Трг Доситейа Обрадовича, 6, 21000, Нови Сад, Сербия; р.т.: +381-21-485-2533.

**Острогонац Стеван** — старший научный сотрудник, AlfaNum – Speech Technologies Ltd, разработчик программного обеспечения, AlfaNum – Speech Technologies Ltd. Область научных интересов: синтез речи, автоматическое распознавание речи, обработка естественного языка, диалоговая система, разработка речевых и языковых ресурсов, машинное обучение, нейронные сети. Число научных публикаций — 18. ostrogonac.stevan@alfanum.co.rs; бул. Войводе Степе, 40, 21000, Нови Сад, Сербия; р.т.: +381-64-845-5302.

**Сузич Синиша** — научный сотрудник лаборатории акустики и речи факультета технических наук, Нови-Садский университет. Область научных интересов: синтез выразительной речи, обработка цифровых сигналов, диалоговая система, машинное обучение, глубокие нейронные сети. Число научных публикаций — 19. sinisa.suzic@uns.ac.rs; Трг Доситейа Обрадовича, 6, 21000, Нови Сад, Сербия; р.т.: +381-21-485-2521.

**Пекар Дарко** — младший научный сотрудник департамента энергетики, электроники и телекоммуникационного инжиниринга факультета технических наук, Нови-Садский университет, главный исполнительный директор, AlfaNum Speech Technologies. Область научных интересов: человеко-машинное взаимодействие, распознавание и синтез речи, идентификация диктора, морфинг речи, статистический анализ, искусственный интеллект. Число научных публикаций — 100. darko.pekar@alfanum.co.rs; бул. Войводе Степе, 40, 21000, Нови Сад, Сербия; р.т.: +381-21-485-2521.

**Поддержка исследований.** Работа выполнена при финансовой поддержке Министерства образования, науки и технологического развития Республики Сербия (проекты TR32035 и OI178027). Авторы также благодарны компании Speech Morphing Inc., г. Кэмпбелл, Калифорния, США, за разрешение использовать свои речевые базы данных для этого исследования.

## Литература

1. *Dall R., Yamagishi J., King S.* Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation // Proceedings of Speech Prosody. 2014. 5 p.
2. *King S., Karaiskos V.* The Blizzard Challenge 2016 // Blizzard Challenge Workshop. 2016. 17 p.
3. *King S., Wihlborg L., Guo W.* The Blizzard Challenge 2017 // Blizzard Challenge Workshop. 2017. 17 p.
4. *Tatham M., Morton K.* Developments in Speech Synthesis // John Wiley & Sons. 2005. 280 p.
5. *Sluijter A. et al.* Evaluation of speech synthesis systems for Dutch in telecommunication applications // Proceedings of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis. 1998. 6 p.
6. *Berg M.* Modelling of Natural Dialogues in the Context of Speech-based Information and Control Systems // PhD Thesis. University of Kiel. 2014. 250 p.
7. *Trouvain J.* Laughing, Breathing, Clicking - The Prosody of Nonverbal Vocalisations // Proceedings of Speech Prosody. 2014. pp. 598–602.
8. *Dall R. et al.* Investigating Automatic & Human Filled Pause Insertion for Speech Synthesis // Proceedings of the Annual Conference of the ISCA. 2014. 5 p.
9. *Székely É., Mendelson J., Gustafson J.* Synthesising Uncertainty: The Interplay of Vocal Effort and Hesitation Disfluencies // 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017). 2017. vol. 2017. pp. 804–808.
10. *Beckman M.E.* Stress and Non-Stress Accent // Foris Publications. 1986. 241 p.
11. *Silverman K. et al.* ToBI: A standard for labeling English prosody // Proceedings of the 2nd International Conference on Spoken Language Processing. 1992. 4 p.
12. *Beckman M.E., Hirschberg J., Shattuck-Hufnagel S.* The original ToBI system and the evolution of the ToBI framework // Prosodic typology: The phonology of intonation and phrasing. 2006. 37 p.
13. *Black A.W., Hunt A.J.* Generating F0 contours from ToBI labels using linear regression // Proceedings of ICSLP. 1996. 4 p.
14. *Wightman C.W.* ToBI or not ToBI // Proceedings of the International Conference on Speech Prosody 2002. 2002. 5 p.
15. *Syrdal A., Hirschberg J., McGory J., Beckman M.* Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody // Speech communication. 2001. vol. 33. no. 1-2. pp. 135–151.

16. *Syrdal A., McGorg J.* Inter-Transcriber Reliability of ToBI Prosodic Labeling // Proceedings of the International Conference on Spoken Language Processing (ICSLP). 2000. 4 p.
17. *Niemann H. et al.* Prosodic processing and its use in Verbmobil // 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97). 1997. vol. 1. pp. 75–78.
18. *Pierrehumbert J., Hirschberg J.B.* The meaning of intonational contours in the interpretation of discourse // Intentions in communication. 1990. pp. 271–311.
19. *Hamza W. et al.* The IBM Expressive Speech Synthesis System // Proceedings of the Eighth International Conference on Spoken Language Processing (ISCLP). 2004. 4 p.
20. *Ze H., Senior A., Schuster M.* Statistical parametric speech synthesis using deep neural networks // 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. pp. 7962–7966.
21. *Delić T., Sečujski M., Suzić S.* A review of Serbian parametric speech synthesis based on deep neural networks // Telfor Journal. 2017. vol. 9. no. 1. pp. 32–37.
22. *Wu Z., Watts O., King S.* Merlin: An Open Source Neural Network Speech Synthesis System // Proceedings of the 9th ISCA Speech Synthesis Workshop. 2016. 6 p.
23. *Seide F., Agarwal A.* Cntk: Microsoft's open-source deep-learning toolkit // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 2135–2135.
24. *Morise M., Yokomori F., Ozawa K.* WORLD: a vocoder-based high-quality speech synthesis system for real-time applications // IEICE Transactions on Information and Systems. 2016. vol. 99. no. 7. pp. 1877–1884.
25. *Tokuda K. et al.* Speech parameter generation algorithms for HMM-based speech synthesis // Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00). 2000. vol. 3. pp. 1315–1318.
26. *Gođevac S.* Transcribing Serbo-Croatian Intonation // Prosodic Typology: The Phonology of Intonation and Phrasing. 2005. 26 p.