Improving HMM Speech Synthesis of Interrogative Sentences by Pitch Track Transformations

Péter Nagy, Géza Németh

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Magyar tudósok krt. 2, Budapest, 1117 Hungary
{nagyp, nemeth}@tmit.bme.hu

Abstract Modeling interrogative sentence prosody is a challenging task due to the significant variation of questions. Prosody is produced by intonation, intensity and duration features. Intonation clearly identifies the type of question in most European languages. If only limited training data is available from certain sentence types, synthetic intonation often lacks accuracy, richness and detail due to averaging, inherent in statistical approaches. In this paper we discuss two rule-based solutions to improve intonation of interrogative sentences. The first approach utilizes a pitch prediction algorithm where a rule-set forms the pitch pattern. The output pattern is combined with an HMM generated F_0 contour and the resulting pattern is used for speech synthesis. Our second solution uses key points to define a scaling function. The position of the key points is described by a rule-set, and the value at intermediate points is calculated in training time in a data-driven way in order to maintain the characteristics of the speakers' voice. The proposed two hybrid rule-based-HMM systems were evaluated by speech experts and by perceptual tests. Our evaluation shows that both approaches could significantly improve the prosodic representation of questions in our framework without deteriorating the perceived naturalness of synthetic speech. The pitch contours generated by the systems were compared and evaluated according to how well they could reproduce the unique characteristics of different Hungarian interrogative sentence types. The analysis shows that both of the methods could successfully model the required patterns. In a separate perception test the interrogative prosody identification rate of the solutions was also measured. The results demonstrate that applying our proposed methods the identification rate of questions could be improved significantly. Although our work focuses on our mother tongue -Hungarian- the methodology can be extended to other languages.

Keywords text-to-speech, prosody modeling, interrogative sentences, HMM speech synthesis

1 Introduction

Intelligible, natural sounding, high quality synthetic speech can be generated with the latest text-to-speech (TTS) systems. During the last few decades several techniques have been proposed including articulatory, formant, diphone concatenation, unit selection and statistical parametric synthesis (Mermelstein, 1973; Charpentier and Stella, 1986; Klatt and Klatt, 1990; Möbius, 2000; Zen et al., 2009). State-of-the-art methods are unit selection (Hunt and Black, 1996; Black and Campbell, 1995) and statistical parametric methods, the latter using mainly Hidden Markov-models (HMM) (Zen et al., 2009; Tokuda et al., 2013) and deep neural networks (DNNs) (Zen et al., 2013; Ling et al., 2013). Unit-selection can produce speech that sounds more similar to the target speaker (Karaïskos et al., 2008) than HMM-based speech synthesis. However, it was found that statistical parametric speech synthesis systems can be as intelligible as natural speech (Yamagishi et al. 2008). In these systems

prosody is modeled either by rules or statistically. The prosody (change in the melody, speed of articulation, rhythm, stress pattern, intensity and tonality) gives expressiveness to speech; it can also reflect the modality of sentences.

In general, modeling prosody in statistical parametric synthesis systems is performed by extracting the descriptive parameters from a training set and by training generative probabilistic models. The synthetic speech is produced by generating parameter streams from the statistical models according to the input text. This method has the advantage that it can create prosody with a high resolution (Zen et al., 2009). By contrast the main drawback of statistical parametric synthesis is that prosody generated from the generative models tends to be over-smoothed. Synthetic prosodic patterns often lack the fine details of natural speech due to inherent averaging of the statistical approach. Statistical averaging improves robustness against sparse data during the model construction phase, but in the synthesis phase this feature hinders the reproduction of detailed characteristics of speech parameters. HMMs can effectively model declarative sentences, but modeling interrogative sentences is still not straightforward because of the rich variability of questions. There is only a limited amount of work in this area (Maia et al., 2003; Gonzalvo et al., 2007; Phan et al., 2013).

To overcome the over-smoothing effect there has been a significant amount of work on prosody modeling in general (mainly for declarative sentences). Some of the solutions utilize rule-based prosody generation solutions (Fordyce, 1998; Atterer, 2002; Zovato et al., 2004), others use data-driven (Langarani et al., 2015) and statistical methods including HMMs (Yamagishi et al., 2005; Yu et al., 2013; Li et al., 2015) and DNNs (Fernandez et al., 2014; Reddy and Rao, 2015). One of the main components of prosody is the pitch value and its pattern over time, and there is a wide range of approaches to predict this pattern. To mention a few, Pan et al. (2000) predicted sentence intonation (together with syllable duration and intensity) utilizing regression analysis; a recurrent neural network (RNN)-based prosody synthesis method was used by Chen et al. (1998) to predict pitch contours. Wu et al. (2001), Tao (2004) and Németh et al. (2007) introduced template-based prosody modeling methods. Suni et al. (2013) uses wavelets to decompose the pitch contour into five temporal scales ranging from microprosody to the utterance level. The different levels of wavelets represent the hierarchical structure of intonation. Likewise a hierarchical approach for modeling and generating pitch in HMM-based speech synthesis of Mandarin speech is introduced in Hsia et al. (2010). The study uses supervised and unsupervised classification and regression trees to predict the prosodic structures. Huang et al. (2013) also proposes a hierarchical pitch pattern retrieval method to generate personalized speech with a more natural pitch contour. The study introduces a modified Fujisaki model, where the pitch contour is generated by hierarchical prosodic commands. The result is then combined with a HMM-based TTS system, and according to the evaluation it produces a more natural synthetic speech than the conventional HMM-based method. The study of Maeno et al. (2014) aims at enhancing prosodic variation of HMM-based expressive speech synthesis. They propose an unsupervised labeling technique to capture phrase-level prosodic variations. The additional contexts are automatically determined by classification of the average difference in the prosodic features of original and synthetic utterances. Gonzalvo et al. (2007) used case based reasoning (CBR) as F_0 estimator for HMM-based speech synthesis in order to improve the prosody estimation. The results of the study show that the CBR-based F_0 estimation is capable of improving performance when non-declarative short sentences are synthesized or reduced contextual information is available.

Besides prosody prediction several approaches are available in the literature for prosody modification or post-processing. Some of these approaches, like pitch synchronous overlap and add

(PSOLA) (Moulines and Laroche, 1995) operate directly on the speech waveform in order to incorporate the desired modifications. Basically PSOLA and its derivatives consist of pitch synchronous analysis segment derivation using pitch markers, determined by pitch extraction algorithms. The length of the analysis window is typically two or four pitch periods. By changing the time intervals between the extracted pitch markers manipulation of pitch can be achieved. Tanaka and Abe (1997) proposed a rule-based fundamental frequency modification algorithm to overcome the naturalness degradation if the pitch is strongly modified. The solution is based on a codebook mapping formulation, and transformation rules are generated using speech data uttered in a different F_0 range. Rentzos et al. (2004) employed a PSOLA based method for pitch transformation in a voice conversion scenario. They used HMMs for estimation of the statistical distribution of spectral and temporal voice correlates of the source and the target speakers. They fed the excitation signal of the source speaker to a TD-PSOLA algorithm for modification of the pitch, speaking rate and intonation features.

In the harmonics plus noise model (HNM) the speech signal is represented in a parametric form as a time-varying harmonic component with a modulated noise component (Laroche et al., 1993). The decomposition allows a flexible modification of the prosody parameters of the speech signal. However, the estimation of the parameters are computationally extensive and the method requires an additional post-processing step to reduce the phase mismatch between frames from different acoustic units (Stylianou, 2001). In case of the discrete cosine transform (DCT) based approaches the pitch modification is performed in the residual domain (Shankar et al., 2001). From the speech signal first the linear prediction (LP) residual is extracted, then pitch markers are computed based on the autocorrelation function. In order to modify the pitch pattern the residual is interpolated using DCT and inverse DCT. Rao and Yegnanarayana (2006) proposes a method for prosody modification that is using instants of significant excitation. The residual signal in the LP analysis is used as an excitation signal. The residual signal is manipulated by using a resampler either for increasing or decreasing the number of samples required for the desired prosody modification. The study found that the manipulation introduces less distortion in the speech signal synthesized using the generated LP residual and LP coefficients.

Only a limited amount of work is available aiming at improving the prosody of interrogative sentences. The study of Gonzalvo et al. (2007) was already discussed previously. This study improves prosody modeling of Spanish speech within the HMM framework using CBR as F_0 estimator. The authors evaluated the system using various phrase types, including interrogatives. Objective measures show that the proposed system could outperform the conventional HMM-based solution, however for longer sentences the difference between the approaches decreases. Subjective evaluation shows that both methods were equally preferred, but the CBR-based estimator could slightly outperform the conventional method in case of interrogatives. For Chinese Wang et al. (2011) analyzed interrogative intonation and its HMM-based synthesis. The authors analyzed the acoustic features of interrogative intonation and their context. Based on the analysis the feature set used for training was expanded to capture the difference between variations of interrogative and declarative intonation. The authors found that with the newly introduced contextual features the perceived prosody of interrogatives in Chinese could be improved. In a subjective perceptual evaluation the test subjects significantly preferred the proposed solution over a baseline system. Deo and Dechpande (2014) aimed at improving the pitch contour modeling for interrogatives in Marathi. In the proposed solution they statistically modeled jitter and shimmer and used these models to modify prosody. The fundamental frequency in the study is converted using linear transformations and also

Gaussian normalization is applied. The performance of the solution was measured in a subjective evaluation and the synthesized utterances were found natural by the listeners. Gubian et al. (2013) proposed a software tool that can translate high level prosodic changes into low level operations and perform them utilizing PSOLA automatically. The authors performed a forced-choice identification evaluation in Italian with the developed software tool. The transformation of declarative sentences to interrogatives proved to be successful since the test subjects judged the original questions and declarative sentences resynthesized as questions not significantly different in terms of naturalness. Mukherjee and Mandal (2014) modifies the HMM generated F_0 contour of interrogative, imperative and exclamatory sentences, based on the Fujisaki model in Bengali. The solution utilizes a two layer supervised classification and regression tree to predict the prosodic word and phrase boundaries. The performed subjective evaluation shows that the proposed method is significantly preferred over the baseline system by the test subjects and also the utterances with transformed F_0 contour were judged more natural.

In this paper we propose solutions to improve the prediction of interrogative sentence prosody in Hungarian HMM-based TTS synthesis. Our goal was to shape the HMM-generated parameter stream to make the system capable of modeling interrogatives better. As a baseline system we used a Hungarian HMM-TTS system (Tóth et al. 2014). Two novel solutions are applied in our proposal in order to modify the baseline HMM generated F_0 stream before vocoding. We use conventional context dependent HMMs to predict a pitch pattern for the input text, then two different parameter modification algorithms are used to improve the pattern. Thus prosody is generated by two-step methods. Instead of the previously presented methods, we selected the rule-based approach for prosody modification. The reason is on one hand that the statistical approach presented in Tóth et al. (2014) could not perform sufficiently enough for interrogatives and on the other hand that Hungarian has a fixed accentuation system. Due to fixed accentuation, syllables with the most important characterizing pitch activity can be determined solely based on textual content and the pitch contour of interrogatives is realized by different speakers with similar patterns. These features allow the rule-based definition of the characterizing sentence stress. We believe that for languages with fixed accentuation systems the rule-based prosody modification method can achieve higher quality and naturalness than other approaches.

One of the drawbacks of the baseline system was that it realized sentence stress on wrong syllables. That could prevent the listener from identifying the synthetic speech as a question. Thus it was important to make our solution able to detect and correct the faulty peaks in the fundamental frequency stream. Examining accentuation in interrogative sentences is one of the most challenging fields in speech research. Stress patterns determine the rhythm and melody of the sentence and they characterize the language. One of the first studies regarding stress detection is the work of Aull and Zue (1985). They extract simple acoustic features from speech, such as duration, energy and F_0 , and then perform a syllable level classification. The syllable is labeled stressed or unstressed. Arons (1994) worked out a solution that relies only on the extracted pitch values and is able to automatically mark accented segments of speech. The algorithm used in the paper adapts to the pitch range of the speaker and automatically selects regions with the highest pitch activity. Xie et al. (2004) use Support Vector Machines (SVMs) as classifiers and compares them to decision tree based classifiers. The study uses the vowel segments from speech, and reports 84.72% detection accuracy for stressed vowels. The method used by Liang (2015) employs HMMs and is based on the concept of one-class classification. The feature vectors are constructed from duration, F_0 , intensity and continuous wavelet transformation-based decomposition of F_0 contours.

In our approach, besides detecting sentence stress and marking the stressed syllable, it was also important to recognize if the sentence stress was realized on the wrong syllable. In order to produce an intonation pattern that lacks the wrong emphases and can serve as an input for our intonation prediction methods, the surplus stressed parts had to be eliminated from the pattern. Two solutions were developed that can utilize the intermediate pitch pattern prepared by the stress detection and correction algorithm. The first uses a rule-based interrogative sentence stress prediction algorithm whose output is implanted into the intermediate pitch pattern. This hybrid method is able to model the intonation of simple (consisting of one clause) interrogatives in Hungarian, and was fine-tuned by a linguist-specialist to produce the most natural pattern. This solution can be generally applied to any speaker's voice, without training a model or without preparing rule-sets specifically tailored to the individual. However, the generalization capability can degrade speaker similarity since sentence stress (e.g. pitch peak heights) is characterized by unique features of the target speaker.

If we do not mimic the question realization style of the speaker we can lose an important prosodic feature. To overcome this issue a speaker dependent rule-based solution was developed. Distinctive key points in the intonation pattern are defined and described with rules. At these distinctive points the exact pitch value is determined in a data-driven way. During HMM training key characteristics of the target speaker's stress realizations are extracted and later used in synthesis time to shape the pitch track. The output generated by this method cannot be used as an intonation pattern that is ready to be fed to the vocoder, but it supports a weighting function that can be applied to the stress removed intermediate pitch pattern. We found that this method can produce a more natural interrogative prosody without losing the high resolution of prosodic features or without degrading speaker similarity. To evaluate the performance of the novel solutions an expert evaluation of the generated pitch contours and subjective perceptual evaluations were carried out. During the evaluation process both of our solutions were compared to the completely statistical baseline solution as well.

The rest of the paper is organized as follows. Section 2 describes the characteristics of prosodic features of Hungarian interrogative sentences focusing on pitch patterns. Section 3 introduces our previous Hungarian HMM-based TTS system that was already tuned to model interrogative sentences. This system is the baseline for our comparisons. Section 4 introduces the sentence stress detection and removal algorithm and the new idea of rule-based prosodic modification of HMM-generated intonation patterns. Section 5 describes the details of the evaluation process and presents the results of both the expert analysis and the perceptual evaluations. Section 6 summarizes the findings and directions of future work.

2 Prosody of Hungarian interrogative sentences

Supra-segmental features are considered as a combination of intonation, stress and rhythm of speech. Statistical investigations proved that these prosodic elements are not universal, but language and text type dependent (Fackrell et al., 2000). The supra-segmental feature set allows the speaker to project higher level structures onto basic segments (phone series) of speech (Németh and Olaszy, 2010). We can express semantic and syntactic information, emotions and even different modalities using diversified prosodic elements. The basic prosodic elements of the supra-segmental feature set for Hungarian (along with other languages) are already discovered (Varga, 1994; Hirst and diCristo, 1998; Olaszy, 2002). Modeling the rich variability of these prosodic elements is a challenging task, but

there are general data-driven (Németh et al., 2007) and rule-based algorithms (Olaszy et al., 2001) to generate these features for Hungarian.

In this paper we focus on enhancing the modeling of interrogative sentences. Fundamental frequency and its pattern over time is a basic prosodic feature that can distinguish different modalities. Generally in Hungarian the placement of sentence stress implies the modality and defines the type of the interrogative sentence. At the sentence stress pitch and intensity emerge from their surroundings and form a contrast. According to Fónagy (1958) the stress may – but not always will – influence the pitch pattern. Hunyadi (1995) says that the stress can be characterized by the complex relation of intensity and pitch. The ratio between pitch and intensity is highly dependent on the language, on the speaker and also on the content of the message. Olaszy et al. (2000) conclude that there is a strong correlation between the modality and the intensity structure. A study that considers mainly F_0 (Olaszy, 2001) found that in case of particularly stressed parts (sentence stress) the pitch pattern is the most distinctive, and it shows a unified behavior. These results suggest that an enhancement in the modeling of pitch patterns of interrogatives in the HMM-based TTS system could improve synthetic prosody and identification of questions.

In everyday speech the most frequently used modality is the declarative form. The intonation pattern of these sentences in Hungarian can be characterized by the term “falling” (Gósy, 2004), i.e. the terminating point of the sentence is always lower in pitch than the beginning point.

The intonation pattern of Hungarian interrogatives express that we are waiting for an answer (Fónagy and Magdics, 1967). This is presented by realizing a notable difference in the F_0 value on certain syllables. A number of studies examined intonation of interrogatives for different languages, e.g. for Chinese (Yuan et al., 2004), for Castilian Spanish (Face, 2004), for Swedish (House, 2002) and also for Hungarian (Olaszy, 2002; Markó, 2007). It is already verified that in Hungarian the intonation pattern itself in most of the cases is sufficient to distinguish interrogative sentences from declarative ones (Gósy and Terken, 1994).

The pitch pattern of Hungarian interrogative sentences is formed according to syllable level rules. There are five types of interrogatives in Hungarian: wh-interrogative (WH), yes-no interrogative (YN), alternative interrogative (AL), control interrogative (CTR) and unfinished interrogative (UN). The rest of the section presents the basic characteristics of Hungarian question types in detail. For a female voice a summary is depicted in Fig. 1. For each type an arrow is placed on the graph area pointing to the position of the sentence stress, and the syllable that holds the peak is highlighted by boldface text under the graph, together with the English translation of the sentences.

In this paper we investigate simple sentences, i.e. each sentence has one clause and one sentence stress. Besides, there is no introduction part in the sentences that could precede the interrogative part. CTR sentences are used if already known information shall be verified. An example of a CTR sentence is highlighted with boldface in the following sample dialogue: “*What time the bus leaves?*” “*At 3 o’clock.*” “**When?**”. CTR interrogatives are characterized with a similar intonation pattern to WH sentences thus Fig. 1 does not present an example for that.

Fig. 1. Comparison of a) WH b) YN c) AL d) UN questions based on their pitch patterns

In case of WH questions (Fig. 1(a)) there is no distinctive intonation pattern, it has a similar shape to declarative sentences, and can be characterized by the falling term (Deme 1962; Olaszy 2002; Gósy

2004). In WH sentences an interrogative pronoun is always present to express the modality. The intonation peak is on the first syllable of the pronoun. A phonetic study (Olaszy, 2002) proved that the peak has a higher pitch value than the starting point of a declarative sentence. This higher peak and the interrogative pronoun reinforces to the listener that the sentence is interrogative.

The F_0 structure of YN questions is unique, they have an uprising intonation part (Gósy and Terken, 1994; Olaszy, 2002), and the exact shape depends on the syllable count of the sentence. In sentences that have only one syllable the pitch value progressively increases throughout the whole sentence (e.g.: “*Én?*”/”*Me?*”). In case of two syllable sentences (e.g.: “*Nekem?*”/”*For me?*”), the first syllable has a low pitch and in the vowel of the last syllable both the rising and the falling part is realized, thus the pattern is called rising-falling. If the sentence has three or more syllables the pitch peak is on the beginning of the penultimate syllable and a sudden drop is also realized in this syllable. The pitch value on the ultimate syllable shows a slight decrease. The part that precedes the peak shows a lower F_0 characteristic than in declarative sentences. In Fig. 1(b) an example sentence is given with five syllables.

AL questions in Hungarian can be divided into two parts that are separated by the word “*vagy*” (that stands for “*or*”). The two parts can be described by different syllable level rules. Prior to the separating word the pitch is low in the first syllable then elevates to a certain level in the second one and keeps this level for the remaining part until the “*or*” word. After the dividing word the sentence stress is realized on the first syllable, followed by a sharp fall in the second syllable. Fig. 1(c) gives an example for this type.

Fig. 1(d) gives an example for an UN sentence. To show that the thought is unfinished in case of UN questions the intonation pattern has a high pitch value at the end of the sentence. This peak can be characterized by a progressive elevation of pitch value through the last three syllables. The pitch value reaches its highest point at the terminating point of the sentence.

3 The baseline HMM-TTS system

Prosody modeling in HMM-TTS systems is carried out by parallel modeling of duration, excitation and spectral parameters (Yoshimura et al., 1999). Hidden Markov model-based speech synthesis was described and presented in a number of studies (Tokuda et al., 2002; Zen et al., 2007; Tokuda et al., 2013). Here we only discuss one of the derivatives that serves as the baseline system for our study and comparisons.

Tóth et al. (2014) introduced new contextual features to model interrogative sentences in the HMM-TTS framework. The basic Hungarian HMM-TTS system (Tóth and Németh, 2008; Tóth and Németh, 2010) was a modified version of the Hungarian derivative of HTS (Zen et al. 2009). In the study of Tóth et al. (2014) context features were extended in order to better describe interrogative sentence stress types. First of all a feature was introduced that reflects the modality of the sentence together with a conditional feature that describes the sub-type if the sentence is interrogative. Additional features were also added that can describe the position of the interrogative stress in the sentence. These features are presented in Table 1.

Table 1. Contextual features in (Tóth et al., 2014) to model interrogative prosody

Feature	Values
Modality of sentence	0: declarative 1: interrogative 2: other (exclamatory, imperative, optative)
Type of interrogative sentence	0: Wh-interrogative 1: Yes-no interrogative 2: Alternative interrogative 3: Other type (represents control and unfinished interrogatives) x: The sentence is not interrogative
Has the current syllable interrogative stress?	0: no, 1: yes
Number of syllables from the previous syllable with interrogative sentence stress	Integer or x: there is no interrogative sentence stress before
Number of syllables to the next syllable with interrogative sentence stress	Integer or x: there is no interrogative sentence stress after

In order to model interrogative sentences better in the HMM-TTS framework the composition of the training corpus had to be planned. In case of the baseline system (Tóth et al., 2014) different ratios of declarative and interrogative sentences in the training and adaptation corpora were evaluated. The system with the new contextual features was compared to an earlier system (Tóth and Németh, 2010). In summary the proposed method could significantly increase the perceived quality of interrogative prosody.

However to reach this improvement the ratio of interrogative and declarative sentences in the training and adaptation corpora had to be changed. This means if no recording is available from the target speaker that contains sufficient amount of interrogative sentences the prosody could not be improved in this system. Distinguishing and marking the different interrogative types is a challenging task based on the phonetic transcription only. Therefore in some cases the content could not be labeled correctly that resulted in inappropriate prosody modeling. These mislabeled sentences could not be identified by the listeners as interrogatives, since they had declarative prosody.

4 Proposed methods

In our approach we have combined the high resolution prosody generation ability of an HMM-based system with the specificity of rule-based sentence stress prediction algorithms. Two alternatives are proposed. The first utilizes the stress prediction algorithm of Profivox (Olaszy et al., 2000), a Hungarian unit concatenation based synthesis system. The output of the rule-based stress prediction is used to modify the F_0 parameter stream generated by the baseline system. The other solution is a novel, speaker dependent rule-based algorithm, where the rules are dynamically constructed according to the speaker's voice characteristics. Since a static, predefined rule-set cannot mimic the target speakers individual stress realization, data loss is present in the first system that could decrease speaker similarity and naturalness. Taking into account the speaker's unique characteristics in the second solution - with dynamic rule generation - the problem of generalization inherent in the first method is overcome.

These solutions do not generate the whole intonation pattern but they shape the HMM generated pitch stream before vocoding. To train the average voice model of the baseline HMM-TTS system speech corpora by eight speakers (4 males and 4 females) were used. The corpus is the Hungarian PPSD speech database (Olaszy, 2013). The corpus contains 1620 declarative and 390 interrogative utterances per speaker. The database is sampled at the rate of 44.1 kHz on 16 bits and windowed by a 25 ms Hanning-window with 5 ms shift. During HMM training for each processed utterance precise labeling is provided together with word and phoneme boundaries. The corpora were annotated and segmented by using HMM forced alignment (Sárosi et al., 2012); thereafter the resulting data were corrected with manual methods based on visual and auditory inspection by specialists. The voice talents misread certain sentences that lead to incorrect forced alignment due to the method fitted the original textual content to the recordings. In these cases the textual contents were modified manually according to the recorded audio, and the phoneme boundaries and durations were corrected manually. The feature vectors used for training consist of 39 mel-cepstral coefficients (including the 0th coefficient), $\log(F_0)$, and aperiodicity measures with their dynamic features extracted by STRAIGHT (Kawahara et al., 1999). As adaptation corpus recordings from a female speaker that contained 220 declarative and 420 interrogative sentences were used. The adaptation corpus is not part of the PPSD speech database, but the recordings were made with the same parameters in a professional studio environment. In order to provide reliable results a separate adaptation corpus was used to train a male voice that supports the results achieved with the female voice. The male corpus contains 300 declarative and 400 interrogative sentences. The recordings were made in the same studio environment as the female corpus with the same parameters. The maximum likelihood linear regression (MLLR) (Tamura et al., 1998) adaptation technique was applied.

A sentence stress detection algorithm is introduced in Section 4.1. It detects if the underlying HMM-based system realized sentence stress on the wrong syllables. The detailed description of the two proposed novel methods is presented in Section 4.2 and 4.3.

4.1 Sentence stress detection during the synthesis phase

The HMM-based system that supplies the baseline pitch pattern may not realize sentence stress on appropriate syllables. Therefore an algorithm had to be developed that is able to detect stressed syllables at inappropriate locations. According to the previously mentioned studies regarding stress detection the central vowel of a syllable can be used as an indicator of the syllable's stress (Domokos et al., 2014). Our goal was to detect sentence stress. This task is simpler than detecting pitch accents at word, syllable and vowel level. It was important to automatically adapt the method to the speaker's pitch range and detect sentence stress with a high accuracy. As a result an intermediate pitch pattern is generated where the inappropriate stressed parts are eliminated. That can be fed to the rule-based prediction methods which transform the pitch track to the appropriate form of the sentence stress that corresponds to the grammatical rules of the language.

The method uses fundamental frequency as a feature and aims to detect improper intonation in interrogative sentences. Significant activity and variability was searched in F_0 . The findings of Arons (1994) were taken into account so the "number of frames above a threshold" measure helped to detect sentence stress. It was necessary to adaptively determine the threshold for the target speaker, thus in the training phase a histogram of pitch values was built. The threshold was chosen to select the top 1% of the extracted values in the distribution. The threshold is set empirically. This stress detection algorithm was tested on 100 randomly selected HMM generated F_0 patterns in order

to tune the threshold measure. Setting the threshold to the top 1% of the distribution resulted in successful marking of the required stressed syllables in our test set in 94 %.

An example for the histogram and the top 1% threshold is given in Fig. 2 for a female speaker's voice that was used as adaptation corpus in the framework. In case of this speaker the calculated 1% threshold was at 229 Hz, the mode of the histogram is at 176 Hz. The vertical axis of the figure is given as a percentage of all frames extracted from the used corpus.

Fig. 2. Pitch histogram for the used adaptation corpus from a female speaker. The calculated 1% threshold is marked in the figure at 229 Hz.

In the synthesis phase each sentence is first evaluated with the above mentioned method. The processing is performed syllable by syllable. If the algorithm locates pitch values inside the syllable that exceed the calculated threshold it marks the syllable stressed. To mark the syllable stressed 5 continuous pitch values should exceed the threshold. Applying the rules described in Section 2 the algorithm first calculates the expected location of stressed syllables and then marks all the syllables as stressed or unstressed. If the algorithm marks a syllable stressed at a wrong position it scales the F_0 values on this syllable to an average. Fig. 3 gives an example for the result of the sentence stress detection and removal in case of a YN interrogative sentence for a female voice.

Fig. 3. Improper pitch peak detection and removal for a YN interrogative. Solid curve: the pattern generated by the baseline system. Dashed curve: the stress-removed intermediate state. Dashed horizontal line: threshold used for stress marking

Considering YN questions in Hungarian (longer than 2 syllables) the pitch peak is on the penultimate syllable of the sentence. We can see that the example in Fig. 3 does not produce this behavior.

Our algorithm works as follows. First we have to determine the type of the sentence, and in case of interrogatives the exact sub-group is defined based on the textual content. The example utterance belongs to the YN group, thus the algorithm looks for the sentence stress in the penultimate syllable. The location of this syllable is determined by the segmentation information provided with the sentence. The algorithm performs a global labeling throughout the whole sentence marking all of the syllables as stressed or unstressed. If the algorithm locates a stressed syllable at an incorrect position (in Fig. 3: any other syllable than the penultimate one), it marks it for further processing. In Fig. 3 we can see that the algorithm marked the first 4 syllables incorrect as the pitch values exceed the calculated threshold. The threshold used for the stressed/unstressed decision is given in the figure with a horizontal dashed line. Then the algorithm scales the pitch value of the marked syllables. An individual scaling factor is calculated for each syllable. First the algorithm calculates the average pitch value in the target syllable. Our goal is to scale this average value throughout the marked syllables corresponding to the mode of the histogram built in the training phase. This mode represents a global unaccented average for the target speaker. Therefore the calculated scaling factor is inversely proportional to the average F_0 value in the syllables. For the example given in Fig. 3 the calculated scaling factors are given in Table 2. It can be seen that the algorithm assigns only one scaling factor to each marked syllable, which results in the average F_0 value in the syllable being equal with the determined mode.

Table 2. Calculated scaling factors for the given example sentence

Syllable number	Original average F_0 value in syllable	Mode of target speakers histogram	Calculated scaling factor
1	224.6 Hz	176 Hz	0.783
2	226.1 Hz	176 Hz	0.778
3	211.4 Hz	176 Hz	0.833
4	204.7 Hz	176 Hz	0.86

4.2 Rule based interrogative sentence stress prediction

Olaszy et al. (2001) introduced a complex rule set to describe the prosodic features (intonation and intensity pattern, phone durations) of sentences based on the textual representation. The components are produced with a three-step superpositional procedure. Since in our study we consider the F_0 model as the main characteristic feature of interrogative sentences, only the procedure of intonation pattern generation is described. Further details can be found in Olaszy et al. (2001). F_0 modeling starts on the highest (supra-segmental) level to produce the intonation pattern on sentence level. This pattern can be described with one of the three terms: rising, uniform and falling. The second step projects the word and syllable level rules to the previously generated sentence pattern. The study defined two structures on word level: neutral word and words with an emphasis. There are three pitch patterns defined on syllable level: rising, leaning and rising-leaning. These rules generate the stress pattern and other syllable level changes in intonation. The final level of processing happens on the segmental level, and here the different micro-intonational changes are built into the F_0 contour. The model was tested and later built into a Hungarian diphone based speech synthesis system: Profivox (Olaszy et al., 2000).

This prosody module was integrated into the synthesis step of our baseline Hungarian HMM-based synthesizer. The changes in the block diagram of the synthesis part compared to the traditional synthesis step (Zen et al., 2009) are depicted in Fig. 4.

Fig. 4. Rule-based prosodic module (with grey background) in the proposed HMM framework

As we can see the spectral and excitation parameters are generated from the context dependent HMMs in the traditional way. But instead of feeding these parameters directly to the synthesis filter, the excitation parameters are first refined by the prosodic module. The input features for the prosody module are the generated pitch stream and the input text. The module first creates an intermediate descriptive form: the so-called intonation matrix. This matrix describes the prosodic features according to the previously mentioned complex rule set. This approach generates prosodic features for two groups of questions. The first group is WH interrogatives where the interrogative pronoun unambiguously defines the type, and the other group is the sum of the rest of the types. The sentence belongs to the second group if no interrogative pronoun can be found in it, and the prosodic parameters are predicted according to the rules of YN sentences. An example for a part of an intonation matrix is given in Table 3.

In each row of this matrix the parameters of a single sound can be found, among others: the phoneme associated with the sound, the height of pitch expressed as a percentage of a reference value, the location of the pitch value expressed as a percentage of the length of the sound. These parameters are used for the prediction of the intonation pattern. Other columns of the matrix give phoneme, SAMPA coded phoneme, intensity and duration related parameters.

Table 3. Part of an intonation matrix generated by the prosody module

<Profivox phoneme code>	<Pitch value in %>	<Pitch location in %>	<Intensity value in %>	<Intensity location in %>	<Phoneme>	<SAMPA code>	<Duration in ms>
...	...	...	...	...	...	...	...
<3>	<81>	<0>	<80>	<60>	<a>	<O>	<84>
<11>	<89>	<100>	<81>	<100>			<56>
<33>	<0>	<100>	<90>	<100>	<r>	<r>	<55>
<7>	<96>	<100>	<90>	<100>	<i>	<i>	<68>
<20>	<0>	<100>	<90>	<100>	<n>	<n>	<62>
<15>	<0>	<100>	<90>	<100>	<g>	<g>	<52>
<2>	<87>	<100>	<99>	<100>	<A:>	<a:>	<160>
<26>	<86>	<100>	<180>	<100>	<z>	<z>	<116>
...	...	...	...	...	...	...	...

In the matrix the 6th column gives the associated phoneme, the 2nd column gives the pitch value in percentage (0 = no predicted change, but interpolation to the next given value) and the 3rd gives its location inside the sound in percentage (100% means the end of the sound). One can describe the intonation pattern with a collection of cut-off points with pitch height and location values. In this intonation pattern one can define pitch values for unvoiced or mixed excitation regions as well in order to make the pattern more precise.

The intonation matrix only gives predicted duration values thus the HMM generated parameter streams have to be synchronized with the baseline system generated intonation pattern. As synchronization information the duration features generated in the HMM framework are used. The predicted intonation pattern is automatically generated with the same length as the HMM generated stream. Now that the two different intonation patterns are at the same length it is possible to combine the two contours with weighted averaging. The weighting factors for the two contours are defined on syllable level. The weight of the modified HMM based intonation pattern is higher on syllables where no interrogative stress should appear, and the weight of the intonation matrix based pattern is higher on syllables where the prosody generation algorithm predicts interrogative stress. The result in case of the YN question of Fig. 3 is depicted in Fig. 5.

Fig. 5. Rule based intonation pattern modification. Solid line: the baseline HMM TTS generated F0 contour (BASE). Dashed line: the rule-based transformed contour (RB).

The HMM generated contour is presented with a solid line. The sentence is a YN question, thus it can be seen that the originally generated pattern does not follow the previously described linguistic rules. The transformed pattern is depicted with a dashed line. Before feeding it to the prosody generation module the pitch track is modified with the method described in Section 4.1 in order to remove the stress on the first four syllables. The sentence stress in the final transformed pattern is present on the penultimate syllable, thus the generated contour corresponds to YN interrogatives. In the rest of paper we denote this system with the RB abbreviation.

4.3 Speaker dependent rule based interrogative sentence stress prediction

The RB sentence stress prediction algorithm could successfully improve modeling of the pitch contour of interrogative sentences. One of the advantages of this solution is that it can be applied to any speakers' voice, making the method speaker independent. However this property is also a drawback. The stress strategy is one of the features of the speakers' voice characteristics. The pitch peak at the sentence stress characterizes the speaker. A predefined peak value that is used to all speakers in the framework can degrade speaker similarity. To overcome this issue a speaker dependent rule based solution (SDRB) was developed, where representative points in the intonation

pattern are modeled dynamically. The shape of the intonation pattern is defined by rules, but values at distinctive key points in the pattern are calculated in a data-driven way.

This solution is depicted in Fig. 6. On the block diagram boxes with grey background are new, compared to the traditional HMM (Zen et al., 2009) training and synthesis phase. In the training phase interrogative stress parameters are extracted from the pitch parameters. Pitch values are collected at key points of interrogative sentences from the training corpus. The key points are defined according to the type of the interrogative sentence and syllable count. A detailed description about the defined key points can be found in Table 4. Based on the input text it is a challenging task to differentiate the type of the question without interpreting the content. Pure syntactic features are difficult to apply. Therefore if an interrogative pronoun can be found in the sentence we mark it WH interrogative otherwise we mark it YN interrogative. Considering AL and UN questions this can be considered as an information loss, but if the intonation pattern of YN interrogatives is applied for these types, the listener may be able to recognize the modality and will interpret the content of the sentence correctly. In order to verify that intelligibility is not deteriorated by this solution, a perceptual evaluation was carried out. The results are described in Section 5.3.

Fig. 6. Dynamic rule based intonation shaper in the HMM framework

In case of WH interrogatives key points are defined at the beginning and at the termination of the sentence regardless of the syllable count. If the sentence contains only one syllable the pitch value at 80% of the syllable duration is also extracted. These values can characterize the contour since in natural utterances the pitch value progressively but not linearly increases and reaches its highest value at the end of the only syllable. If there is more than one syllable the measurements are complemented with the highest pitch value in the first syllable, with the value at the end of the first syllable and with the terminating value of the second syllable. Thus in this case five key points are defined. These values can characterize the intonation pattern of WH interrogatives with more than one syllable since in these questions only the sentence stress represents a peak in the pitch pattern. There are no other emphasized parts. Then the average of the pitch values is calculated individually

for each key point. In Section 4.1 our sentence stress detection algorithm was described that included a step for building a histogram of pitch values. Table 4 summarizes the locations of the extracted key points and gives examples for all types in case of a female speaker.

For YN interrogatives the starting and ending points of the sentence are considered key points again, but here we have to handle separately the one, two and more than two syllable cases. If there is only one syllable the pitch value is increasing progressively in a nonlinear way until the terminating point, so the value at 80% of the syllable duration is also extracted. As we can see the character of the intonation pattern of one syllable YN and WH interrogatives is the same. The only difference is that in case of WH interrogatives the pitch peak is higher. If the utterance contains two syllables the sentence stress is on the second syllable. Therefore the values at the start and at the highest peak in the second syllable are also extracted. In case of three or more syllables key points are defined at the beginning of the penultimate syllable, at the highest value of the penultimate syllable and at the terminating point of the penultimate syllable. For YN interrogatives with different syllable count Table 4 gives the location of the extracted key points and provides examples used in the study.

Table 4. Extracted F0 averages at key points and normalized stored values for the target speaker for different interrogative types

Interrogative type	Syllable number	Key point number	Key point location [in % of syllable length]	Extracted average	Normalizing factor	Stored key point value
WH	1	1	0% of syllable	165.5 Hz	176 Hz	0.94
		2	80% of syllable	184.8 Hz		1.05
		3	100% of syllable	225.3 Hz		1.28
	>1	1	0% of first syllable	183.1 Hz	176 Hz	1.04
		2	Highest value of first syllable	221.8 Hz		1.26
		3	100% of first syllable	184.8 Hz		1.05
		4	100% of second syllable	172.5 Hz		0.98
		5	100% of last syllable	160.2 Hz		0.91
YN	1	1	0% of syllable	158.4 Hz	176 Hz	0.9
		2	80% of syllable	174.3 Hz		0.99
		3	100% of syllable	209.5 Hz		1.19
	2	1	0% of first syllable	158.4 Hz	176 Hz	0.9
		2	0% of second syllable	176 Hz		1
		3	Highest value of second syllable	213 Hz		1.21
		4	100% of second syllable	165.5 Hz		0.94
	>2	1	0% of first syllable	149.6 Hz	176 Hz	0.85
		2	0% of penultimate syllable	161.9 Hz		0.92
		3	Highest value of penultimate syllable	211.2 Hz		1.2
		4	100% of penultimate syllable	197.1 Hz		1.12
		5	100% of last syllable	160.1 Hz		0.91

The “Extracted average” column of Table 4 contains the averaged values at the key points for the target speaker of our study. These values are extracted during the training phase, together with the mode of the F_0 histogram. This mode is used to normalize the averages at key points and results in the dimensionless stored values given in the “Stored key point value” column.

In the synthesis phase the type of the question is determined according to the input text and the corresponding rule based pattern is selected according to the syllable count of the utterance. The pattern is constructed according to rules, but the values at key points are determined by the previously described data-driven approach. Between the predefined key points the generated pattern is linearly interpolated. The resulting curve can be used as a weighting function on the HMM generated contour. Since the location of the key points are described with syllable level rules, time warping of the produced weighting functions is necessary using the HMM generated duration constraints. Examples of such time warped weighting functions are given in Fig. 7 for the different question types and syllable variations for concrete realizations. In all cases the location of the defined key points are given with black dots in the figures.

Fig. 7. Comparison of scaling factors for a) one syllable WH b) multiple syllable WH c) one syllable YN d) two syllable YN e) three or more syllable YN interrogative sentences. The key points are represented as dots.

In Fig. 7(d) an initial fall in the scaling factor can be observed. This example starts with a pause thus it is not part of any syllable. The scaling factor is interpreted on syllables only, and the default value for pauses is 1. The fall occurs because the algorithm defines an initial value for the start of the first syllable, which does not coincide with the beginning of the sentence. Between the beginning of the first syllable and the sentence, the values are linearly interpolated that results in the initial drop. The resulting F₀ contour in case of the YN interrogative - used as an example in previous sections - is depicted in Fig. 8. The original baseline HMM generated pattern is shown with the solid line, the dashed grey line is the result of the applied weighting and scaling procedure. Compared to Fig. 5 it can be observed that the pitch value at the sentence stress on the penultimate syllable is different and that the fall of the value in the penultimate and final syllable is less steep. Since the change in the F₀ value is not so sudden the synthesized speech sounds more natural because this wider peak provides audible differences compared to the speech generated by the RB pattern.

Fig. 8. Speaker dependent rule based intonation pattern modification. Solid line: baseline HMM TTS generated F₀ contour (BASE). Dashed line: the speaker dependent rule-based transformed contour (SDRB)

5 Evaluation

To evaluate the performance of our solutions the generated pitch contours were inspected by experts and listening tests were also carried out. The comparison of pitch contours evaluates how well our novel solutions (RB and SDRB) could reproduce the proper interrogative patterns. The generated contours are compared against the baseline HMM solution (BASE). It is also important to evaluate if the quality of the generated sentences is consistent so objective measurements were carried out. A separate perceptual test was designed about the identification of the samples as interrogatives.

5.1 Pitch contours

Illustrative examples for different types of interrogative pitch contours are shown in Fig. 9-Fig. 11. On the figures the curve drawn with a thin solid line belongs to the BASE intonation pattern. The dashed line represents the pitch track generated by RB system. The SDRB solution with scaling functions is drawn with the thick solid line. The textual representation is also given on the figures as syllables together with vertical lines that represent the word boundaries. The test sentences used during the evaluation were not included in the training or the adaptation corpora. All types of interrogative sentences were collected from freely available Hungarian novels. Altogether 450 sentences were collected, and from this corpus 120 sentences were selected where different interrogative types are present (24 sentences/type).

Fig. 9 gives an example of a WH sentence. As described in Section 2 this type of interrogative is characterized by one intonation peak on the first syllable of the interrogative pronoun. In this example “*Mikor*” (“*When*”) is the pronoun. All the three solutions (BASE, RB and SDRB) were able to realize the pitch peak on the first syllable of the interrogative pronoun. This behavior stems from the fact that in Hungarian the declarative intonation pattern is not significantly different from WH patterns, since in case of declarative sentences usually the first syllable carries the sentence stress. However, even if in many cases the BASE solution synthesized the sentences as declaratives, the synthetic prosody was realized according to the proper rules, as the first syllable of the interrogative pronoun was stressed. As we can see on the figure the peak is higher on the first syllable in case of RB and SDRB solutions. Olaszy (2002) found that the only intonation related difference between WH interrogatives and declaratives is the higher peak in pitch on the first syllable. On the penultimate and final syllables of the sentence the RB system realized slight stress. This is a result of the three-step pitch contour generation method described by Olaszy et al. (2002). According to the listening tests described in Section 5.3 it can be considered as a feature that does not prevent the listener from recognizing the sentence as an interrogative. In order to assess the pitch contours 3 speech experts were involved in the evaluation. In case of the BASE solution the experts agreed that 21 from the 24 investigated WH interrogative sentences can be identified as questions. In the remaining three cases two of the experts found the realized pitch peak on the first syllable of the interrogative pronoun too low to classify the utterance as interrogative. Regarding the RB solution the experts agreed that 22 from the investigated 24 can be identified as interrogative. From the remaining two utterances, in one case all of the experts agreed that the realized pitch peak is too low, and in the other case two of the experts judged the realized peak too low. From the 24 WH interrogative utterances synthesized with the SDRB solution all of the experts agreed that all 24 sentences could be identified as interrogatives. According to the experts, comparing the SDRB solution to the BASE and RB solutions, the main feature that improved the recognition of this interrogative type is that SDRB realized the highest pitch peak on the first syllable of the interrogative pronouns. All of the

experts agreed that the example presented in Fig. 9 can be identified as a question with all three solutions.

Fig. 9. Pitch contour of a synthesized WH-interrogative. Thin solid line: BASE, dashed line RB, thick solid line: SDRB pattern

In simple YN questions there is typically one intonation peak and if the sentence has more than 2 syllables this is realized on the penultimate syllable. The pitch increases in the penultimate syllable and falls to a lower value still in this syllable then gradually decreases until the end of the sentence. An example is given in Fig. 10 for a YN interrogative sentence intonation pattern. We can see that the BASE solution could not model this behavior. The BASE intonation pattern seems to be closer to a declarative sentence: the intonation peak is on the first syllable and pitch progressively decreases until the terminating point of the sentence. In case of the RB and SDRB solutions the previously described stress scaling method removed the accentuation of the first syllables. The RB system realized the pitch peak on the penultimate syllable that is followed by a sudden drop and the pitch value for the ultimate syllable remains low. The SDRB method also realized the rising-falling shape on the penultimate syllable, but the pitch value on the final syllable remained higher. This stems from the realization of the solution since it scales the baseline HMM generated intonation pattern, and on the final syllable the BASE system generated high pitch values, so the solution preserved the values generated by the baseline system.

According to the experts only 1 from the 24 utterances synthesized with the BASE method could be identified as a YN interrogative. From the remaining 23 utterances in case of 21 sentences the experts agreed that the utterance cannot be identified as a question. In case of 2 sentences two experts said that the utterance can be identified as an interrogative, but the third expert disagreed because the realized pitch peak on the penultimate syllable was too low. In case of the utterances synthesized with the RB solutions the experts agreed that 17 sentences could be identified as YN interrogatives, and 4 could not be identified as interrogative because of the missing or too low pitch peak on the penultimate syllable. In the remaining 3 cases two experts agreed that they can be identified as questions, but one of the experts disagreed because of the low pitch peak. Considering the utterances synthesized with the SDRB solutions the expert agreed that 21 samples could be identified as YN interrogatives. In the remaining 3 cases two of the experts judged the utterances as interrogatives, the third expert said the realized pitch peak in these is too low to classify the sentence as an interrogative. All of the experts agreed that the sentence presented in Fig. 10 in case of the

BASE system could not be identified as an interrogative, but the sentence in both of the RB and the SDRB cases could be recognized as a question.

Fig. 10. Pitch contour of a synthesized YN interrogative. Thin solid line: BASE, dashed line RB, thick solid line: SDRB pattern.

As described in the previous section we did not model separately other types of interrogatives (AL, UN patterns) in this study. Fig. 11 gives an example about how the RB and the SDRB solutions perform for an AL question. In AL interrogative sentences the prosody is based on syllable level rules. The realization of this type is described in Section 2. As we can see none of the solutions could replicate this behavior. The BASE solution realized a declarative intonation pattern. The rule-based solutions realized a contour that is similar to the contour of YN interrogatives. This was the expected behavior since during the training phase we classified all questions without an interrogative pronoun as YN interrogative.

From the 24 samples synthesized with the BASE solution the experts identified in agreement 9 utterances as interrogatives and they agreed that 13 could not be identified as a question. In the remaining 2 cases two of the experts found that because of the incorrectly realized pitch pattern the sentence cannot be identified as a question, but the last expert said these are also recognizable. In case of the RB solution the experts agreed that 18 utterances can be classified as questions, and 2 utterances cannot, because of the barely audible intonation peak. In case of the remaining 4 samples one of the experts said that they cannot be identified as questions because of the low pitch peak. The other two experts disagreed in these cases. According to them these utterances can still be classified as interrogatives despite the lower pitch value at the peak. From the 24 utterances synthesized with the SDRB solution the experts agreed that 21 can be identified as a question and 1 cannot because of low pitch value at the intonation peak. In case of the last 3 utterances, one of the experts found the realized pitch peak too low, but the other two experts still found the sentences identifiable as questions. All of the experts agreed that in case of the RB and SDRB solutions the realized F_0 contour does not match the grammatical rules of Hungarian spoken language. However they also agreed that synthesizing these type of questions as YN interrogatives could help identify them as questions. To verify this statement AL questions were included in the perceptual evaluations. The experts agreed that the sample provided in Fig. 11 could be recognized as interrogative in case of the RB and SDRB solutions.

Fig. 11. Pitch contour of a synthesized AL question. Thin solid line: BASE, dashed line RB, thick solid line: SDRB pattern.

5.2 Perceptual Evaluation

To measure the subjective preference a listening test was carried out. Three systems were involved in the comparison mean opinion score (CMOS) listening test. The same 120 interrogative sentences were synthesized with each system (BASE, RB, SDRB). These questions were the same that were used during the pitch contour evaluation. The test set contains examples of each question type. By including AL and UN interrogatives in the test set we wanted to evaluate if our novel methods (using YN interrogative intonation pattern on these types) could result in synthetic speech that is appreciated better by the listeners than the BASE solution. From the synthesized test set of the BASE solution 30 sentences were randomly selected, and the same 30 sentences were drawn from the RB and SDRB systems. This means the use of a total of 90 sentences (3 system * 30 sentences) in the subjective evaluation. All the subjects listened to 30 sentence pairs, 10 pairs comparing the baseline HMM-based solution (BASE) with the intonation matrix based solution (RB), 10 pairs comparing HMM with the data-driven rule-based solution (SDRB) and 10 pairs comparing RB with SDRB. The spoken text was the same for each individual pairing. The order of the pairs was pseudo-random to avoid the memory effect. The listeners had to answer the question *"Which sentence's prosody do you perceive closer to natural interrogative prosody?"* on a three point scale (*"The first is closer"*, *"They are the same"*, *"The second is closer"*).

Altogether 31 test subjects (14 female and 17 male) were involved in the test. The test subjects were all native Hungarian speakers without any known hearing loss. The listening test was web based, the average age was 37, the youngest subject was 16, and the oldest was 75 years old. 2 subjects were speech experts.

The result of the test is shown in Fig. 12.

Fig. 12. Results of the CMOS listening test for the female voice

It can be seen that based on the listener's subjective opinion both of the novel solutions outperformed the baseline HMM speech synthesis method. The significance for each system was tested with a Student t-test. The t-test showed significant difference in the BASE-RB and BASE-SDRB comparison cases if $\alpha=0.05$. It can be also observed in Fig. 12 that if we compare SDRB and RB systems the test subjects slightly preferred the SDRB method, but the difference is not significant.

In order to support the applicability of the proposed methods a separate listening test was carried out using a male voice. The same 120 interrogative sentences were synthesized as in case of the previously discussed female voice. During the preparation of the evaluation from the sample set of the BASE solution 30 utterances were randomly selected again, and the same 30 sentences were drawn from the RB and SDRB systems. The format of the evaluation is the same as the one previously described. Meaning a total of 90 sentences (3 system * 30 sentences) were included and all the test subject listened to and evaluated 30 sentence pairs. In order to avoid the memory effect the order of the pairs was pseudo-random. The listeners once again had to answer the question "*Which sentence's prosody do you perceive closer to natural interrogative prosody?*" on a three point scale ("*The first is closer*", "*They are the same*", "*The second is closer*").

21 test subjects (7 female and 14 male) were involved in the test. The test subjects were all native Hungarian speakers without any known hearing loss. The listening test was web based, the average age was 33, the youngest subject was 16, and the oldest was 66 years old. None of the subjects were speech experts. The results of the evaluation is presented in Fig. 13.

Fig. 13. Results of the CMOS listening test for the male voice

It can be seen that for the male voice the listeners judged the interrogative prosody of the novel solutions more natural than the baseline method. For this second evaluation the significance for each

comparison was tested with Student t-test. The test showed significant difference in the BASE-RB and the BASE-SDRB comparison cases if $\alpha=0.05$. It is also worth mentioning that in case of the male voice the listeners preferred the SDRB solution over the baseline slightly more than the RB solution which means that the SDRB method was able to better approach the desired natural interrogative prosody, however this difference is not statistically significant. Fig. 13 compares the SDRB and RB solutions also. For the male voice again the test subjects preferred the utterances synthesized with the SDRB solution, but the difference is not significant.

We also performed a mean opinion score (MOS) test, where the subjects had to evaluate the overall quality of the synthetic sentences (not mentioning interrogative prosody). Listeners had to evaluate naturalness on a scale of 5, where 1 means *“Not understandable”* and 5 means *“Can be confused with human speech”*. The results show (BASE: 3,43; RB: 3,45; SDRB: 3,44) that the applied rule-based intonation shaping did not degrade but rather slightly improved the naturalness of the synthesized speech however the difference is not statistically significant.

5.3 Perceptual identification of interrogative prosody

To measure how well the listeners could identify the test sentences as interrogatives a separate listening test was carried out using the female voice. The same 30 sentences from each system (BASE, RB, SDRB) were used in the evaluation as in the previous subsection. This set was extended with 30 more sentences synthesized with the BASE solution as declarative sentences. This set serves as a control group for the other systems and is presented with the name DECL. During the evaluation the subjects listened to and evaluated 28 sentences, 7 from each system. The order of the sentences was randomized to avoid the memory-effect and the subject could listen to one sentence at a time. The listeners had to answer the question *“How do you judge the modality of the heard sentence?”* on a four point scale (*“It is declarative”*, *“It is rather declarative”*, *“It is rather interrogative”*, *“It is interrogative”*).

Altogether 38 test subjects (12 female and 26 male) were involved in the test. The test subjects were all native Hungarian speakers without any known hearing loss. The listening test was web based, the average age was 46, the youngest subject was 24, and the oldest was 83 years old. None of the subjects were speech experts. The results of the evaluation are presented in Fig. 14.

Fig. 14. Results of the interrogative prosody identification evaluation considering all question types together

Fig. 14 gives the results for the case when all question types are assessed together. The SDRB solution outperforms the others. The significance was tested with two sample t-tests, and it showed that the SDRB sentences were identified significantly better than any other system if $\alpha=0.05$. The RB

system was identified significantly better than the DECL system, but the difference between RB and BASE systems is not statistically significant.

It was also examined how well the test subjects could identify the different types of interrogatives separately. The results for the separate analysis are depicted in Fig. 15.

Fig. 15. Results of the interrogative prosody identification evaluation. a) Considering only YN type b) Considering only WH type c) Considering only AL and UN types

If we assess only the YN interrogative type (Fig. 15(a)), we can see that both RB and SDRB outperformed the BASE and DECL systems significantly ($\alpha=0.05$). In case of SDRB the identification rate is 88%, for RB it is 76%. We could not observe a statistically significant difference between the BASE and the DECL sentences in case of YN questions.

Fig. 15(b) shows the results for the WH type only. In this case each system (BASE, RB, SDRB) outperformed DECL significantly ($\alpha=0.05$). The identification rate for the SDRB solution is over 95%.

However it is important to note that in case of DECL sentences more than 50% of the listeners assessed them as interrogatives. Since WH interrogative and declarative prosody are similar to each other, listeners judged sentences with the interrogative pronoun as interrogatives. Fig. 14 shows that 18% of the DECL sentences were considered as interrogatives by the test subjects, and this is the result of WH interrogatives.

Considering AL and UN interrogatives - Fig. 15(c) - SDRB had significantly higher identification rate than any other system presented in the evaluation. The identification rate is slightly over 73%. Applying the YN pitch pattern in these cases could improve the identification, and most of the test subjects could recognize these sentences as interrogatives, even when the pitch pattern itself is formed contrary to grammatical rules.

6 Conclusions and future work

Due to the rich variation of question types in Hungarian it is a challenging task to model their supra-segmental structure. In this paper we introduced two rule-based solutions and integrated them into the Hungarian HMM-based speech synthesis system. The first solution is based on a speaker independent intonation prediction algorithm that was originally used in a previous Hungarian unit concatenation TTS system. The second method is a novel idea and works with a cut-off point based linearized intonation scaling function in a data-driven manner. The results show that combining the capability of the statistical parametric speech synthesis technique with rule-based approaches could significantly improve the perceived prosody and identification rates of interrogatives in Hungarian. Although these rule-based approaches cannot model all the different types of questions (AL and UN types are handled as YN questions), the generated interrogative prosody for all types of questions results in synthetic speech that could be identified as questions. The results also prove that the applied intonation shaping techniques did not deteriorate the naturalness of synthesized speech.

Although our work focuses on our mother tongue - Hungarian- the methodology can be extended to other languages. Languages may have different accentuation systems, there are languages with fixed (e.g. Hungarian, French) and free (e.g. English, Russian) accentuation. Our solutions can be adapted to other fixed accentuation languages by modifying the underlying rule set. To adapt the RB method a complex rule-set has to be defined for a specific language by linguists to accurately predict the intonation pattern. By contrast the adaptation of our SDRB method to new languages may be simpler. The number and position of key points used by the algorithm can be easily modified inside the framework. Only these parameters have to be adjusted to make the solution able to work with other languages with a fixed accentuation system.

There might be different ways to improve interrogative sentence prosody prediction further. As we could observe at the evaluation of the pitch contour of the YN interrogative question in Section 5.1 our sentence stress detection algorithm could successfully remove the faulty generated pitch peaks in case of the test sentences based on the empirically set threshold. However other incorrect emphasized prosodic parts may be present in the sentence. These peaks have to be removed in order to further improve quality. With the intonation prediction algorithm described in Section 4.3 we could capture the targets speakers' stress strategy with a data driven approach. However, a reasonable solution for further improvement would be statistical modeling of the value at the cut-off points. A suitable statistical approach could replace the fixed values with Gaussian distributions or with Gaussian Mixture Models. We also plan to develop a Deep Neural Network based solution for sentence stress prediction specifically aiming at interrogative sentences. We only dealt with

intonation in our study but other supra-segmental features also include intensity and duration features. Appropriate modeling and transformation of these parameters could also further improve the quality of synthesized interrogatives.

Acknowledgement

Our sincere thanks go to Prof. Gábor Olasz for his encouraging comments and advice. We thank the listeners for participating in the subjective test. The staff of the Speech Technology and Smart Interactions Laboratory contributed with useful comments to this manuscript. We also would like to thank the anonymous reviewers for their valuable suggestions and remarks that helped to improve this manuscript. This research is partially supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech prosody (SNSF no IZ73Z0 152495-1).

References

- Arons, B.: Pitch-based Emphasis Detection For Segmenting Speech Recordings, in Proc. of International Conference on Spoken Language Processing, Vol. 4, pp. 1931-1934 (1994)
- Atterer, M.: Assigning prosodic structure for speech synthesis: a rule-based approach, In Proceedings of the First International Conference on Speech Prosody SP2002, Aix-en-Provence, pp. 147-150 (2002)
- Aull, A., Zue, V. W.: Lexical stress determination and its application to large vocabulary speech recognition, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'85, pp. 1549-1552 (1985)
- Black, A. W. and Campbell, N., Optimizing selection of units from speech database for concatenative synthesis, In Proc. EUROSPEECH-95, pp. 581-584 (1995)
- Charpentier, F., Stella, M.: Diphone synthesis using an overlap-add technique for speech waveforms concatenation, Proc. of ICASSP'86, vol. 11, pp. 2015-2018 (1986)
- Chen, S. H., Hwang, S. H., Wang, Y. R.: An RNN-based prosodic information synthesizer for Mandarin text-to-speech, IEEE Trans. Acoust., Speech, Signal Process., vol. 6, no. 3, pp. 226-269 (1998)
- Deme, L.: A hanglejtés [The Intonation], In Tompa J. The system of Hungarian language II., Akadémiai Kiadó, pp. 503-522 (1967) (in Hungarian)
- Deo, R. S. and Deshpande, P. S.: Pitch contour modelling and modification for expressive Marathi speech synthesis, International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2455-2458 (2014)
- Domokos, J., Stan, A., Giurgiu, M.: An approach to lexical stress detection from transcribed continuous speech using acoustic features, Telecommunications Forum Telfor, pp. 525-528 (2014)
- Face, T. L.: The Intonation of Absolute Interrogatives in Castilian Spanish, Southwest Journal of Linguistics, Vol. 23, No. 2, pp. 65-79 (2004)
- Fackrell, J., Vereecken, H., Buchmann, J., Martens, J., Coile, B.: Prosodic variation with text type, In Proc. ICSLP, vol. 3, pp. 231-234 (2000)
- Fernandez, R., Rendel, A., Ramabhadran, B., Hoory, R.: Prosody Contour Prediction with Long Short-Term Memory, Bi-Directional Deep Recurrent Neural Networks, In Proc. INTERSPEECH 2014, pp. 2268 – 2272 (2014)
- Fordyce, C. S., and Ostendorf, M.: Prosody prediction for speech synthesis using transformational rule-based learning, Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), pp. 682-685 (1998)
- Fónagy I.: A hangsúlyról [About stress], in Nyelvtudományi Értekezések 18. Budapest, Akadémiai Kiadó. (1958) (in Hungarian)

- Fónagy, I., Magdics, K.: A magyar beszéd dallama [The tune of Hungarian speech], Akadémiai Kiadó, p. 317 (1967) (in Hungarian)
- Gonzalvo, X., Iriondo, I., Socoró, J. C., Alías, F., Monzo, C.: HMM-based Spanish speech synthesis using CBR as F0 estimator, ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP07), pp. 7-10 (2007)
- Gonzalvo, X., Socoró, J. C., Iriondo, I., Monzo, C., & Martinez, E.: Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish, In Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6), pp. 362-367 (2007)
- Gósy, M.: Fonetika, a beszéd tudománya [Phonetics, the science of speech], Osiris, p. 350 (2004) (in Hungarian)
- Gósy, M., Terken, J.: Question marking in Hungarian: Timing and height of pitch peaks, Journal of Phonetics, vol. 22., pp. 269-281 (1994)
- Gubian, M., Asano, Y., Asaridou, S., Cangemi, F.: Rapid and smooth pitch contour manipulation, Proceedings of the Workshop on Tools and Resources for the Analysis of Speech Prosody, pp. 31-35 (2013)
- Hirst, D., diCristo, A.: Intonation systems. A Survey of Twenty Languages, Cambridge University Press, pp. 1-432 (1998)
- House, D.: Intonational and visual cues in the perception of interrogative mode in Swedish, In J. H. L. Hansen & B. Pellow (eds), 7th International Conference on Spoken Language Processing (ICSLP), pp.1957-1960 (2002)
- Hsia, C. C., Wu, C. H. and Wu, J. Y.: Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 8, pp. 1994-2003 (2010)
- Huang, Y. C., Wu, C. H. and Lin, S. L.: Personalized natural speech synthesis based on retrieval of pitch patterns using hierarchical Fujisaki model, Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7844-7848 (2013)
- Hunt, A. and Black, A.W., Unit selection in a concatenative speech synthesis system using a large speech database, in Proc. ICASSP-96, pp. 373-376 (1996)
- Hunyadi, L.: Acoustic cues to sentential stress in Hungarian and their measurement, In Hunyadi, L.–Gósy, M.–Olaszy, G. Studies in Applied Linguistics. Debrecen, pp. 29–48 (1995)
- Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C.: The Blizzard Challenge 2008, In Proceeding Blizzard Challenge Workshop 2008 (2008)
- Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Speech Communication, vol. 27, no. 3-4, pp.187-207 (1999)
- Klatt, D. H., Klatt, L. C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers, The Journal of the Acoustical Society of America vol. 87., issue 2, pp. 820-857 (1990)
- Langarani, M. S. E., van Santen, J. P., Mohammadi, S. H., and Kain, A.: Data-driven foot-based intonation generator for text-to-speech synthesis, in Proc. of INTERSPEECH, 2015., pp. 1596-1600 (2015)
- Laroche, J., Stylianou, Y., Moulines, E.: HNS: Speech modification based on a harmonic + noise model, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 550-553 (1993)
- Li, Y., Tao, J., Hirose, K., Xu, X., Lai, W.: Hierarchical stress modeling and generation in mandarin for expressive Text-to-Speech, Speech Communication, Volume 72, September 2015, pp. 59-73 (2015)
- Liang, H.: Detecting emphasised spoken words by considering them prosodic outliers and taking advantage of HMM-based TTS framework, Laboratory TIK, ETH Zurich, Switzerland, Tech. Rep. TIK-Report-362, p. 6 (2015)

- Ling, Z. H., Deng, L., Yu, D., Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis, *IEEE transaction on Audio, Speech, and Language Processing*, Vol. 21, No. 10, pp. 2129-2139 (2013)
- Maeno, Y., Nose, T., Kobayashi, T., Koriyama, T., Ijima, Y., Nakajima, H., Mizuno, H., Yoshioka, O.: Prosodic variation enhancement using unsupervised context labeling for HMM-based expressive speech synthesis, *Speech Communication*, Volume 57, pp.144-154 (2014)
- Maia, R., Zen, H., Tokuda, K., Kitamura, T., Resende Jr., F.G.: Towards the development of a Brazilian Portuguese text-to-speech system based on HMM, *Proc. of Interspeech*, pp. 2465-2468 (2003)
- Markó, A.: Kérdő funkciójú hanglejtésformák a spontán beszédben [Intonation patterns of interrogatives in spontaneous speech], In Gósy M. *Beszéd kutatás 2007*, pp. 59-74 (2007) (in Hungarian)
- Mermelstein, P.: Articulatory model for the study of speech production, *Journal of the Acoustical Society of America* 53 (4), pp. 1070-1082 (1973)
- Moulines, E., Laroche, J.: Non-parametric techniques for pitch-scale and time-scale modification of speech, *Speech Communication*, vol. 16, pp. 147-158 (1995)
- Möbius, B.: Corpus-based speech synthesis: methods and challenges, *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart)*, AIMS 6(4), pp. 87–116 (2000)
- Mukherjee, S., Das Mandal, S.K.: F0 contour generation and synthesis using Bengali HMM-based speech synthesis system, *International Journal of Speech Technology*, vol. 18, pp. 25-36 (2014)
- Németh, G., Olaszy, G.: *A Magyar Beszéd (The Hungarian Speech)*, Akadémiai Kiadó, pp. 171-208 (2010) (in Hungarian)
- Németh, G., Fék, M., Csapó, T. G.: Increasing Prosodic Variability of Text-To-Speech Synthesizers, In *Proc. Interspeech*. Antwerp, Belgium, pp. 474–477 (2007)
- Olaszy, G.: A beszéd akusztikai-fonetikai elemzése és modellezése, különös tekintettel a korszerű beszédépítés követelményeire [Phonetic-Acoustic Analysis and Modeling of Speech, Concerning of up to date Computer based Speech Generation], D.Sc. thesis, Hungarian Academy of Sciences (2001) (in Hungarian)
- Olaszy, G.: The most important prosody patterns of Hungarian, *Acta Linguistica Hungarica*, Vol. 49(3-4). pp. 277-306 (2002)
- Olaszy, G., Precíziós, párhuzamos magyar beszédatadbázis fejlesztése és szolgáltatásai [Development and services of a Hungarian precisely labeled and segmented, parallel speech database] (in Hungarian), *Beszéd kutatás 2013 [Speech Research 2013]*, pp. 261–270 (2013)
- Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C., Gordos, G.: Profivox – a Hungarian TTS System for Telecommunications Applications, *International Journal of Speech Technology*. Vol 3-4., pp. 201-215 (2000)
- Olaszy, G., Németh, G., Olaszi, P.: Automatic Prosody Generation - a Model for Hungarian, *Proc. of Eurospeech 2001*, pp. 525-528 (2001)
- Pan, N. H., Jen, W. T., Yu, S. S., Yu, M. S., Huang, S. Y. and Wu, M. J.: Prosody model in a Mandarin text-to-speech system based on a hierarchical approach, in *Proc. IEEE Int. Conf. Multimedia Expo*, Vol. 1, pp. 448-451 (2000)
- Phan, S. T., Vu, T. T., Duong, C. T., & Luong, M. C.: A study in Vietnamese statistical parametric speech synthesis based on HMM, *International Journal of Advances in Computer Science and Technology*, 2(1)., pp. 1-6 (2013)
- Rao, K. S., Yegnanarayana, B.: Prosody Modification Using Instants of Significant Excitation, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972-980 (2006)
- Reddy, V. R. and Rao, K. S.: Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks, *Neurocomputing*, vol. 171, pp. 1323-1334 (2015)
- Rentzos, D., Vaseghi, S., Yan, Q., Hi, C.-H.: Voice conversion through transformation of spectral and intonation features, in *Proc. ICASSP 2004*, vol. 1, pp. 21-24 (2004)

- Sárosi, G., Tarján, B., Balog, A., Mozsolics, T., Mihajlik, P., & Fegyó, T.: On modeling non-word events in Large Vocabulary Continuous Speech Recognition, In International Conference on Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd, pp. 649-653 (2012)
- Shankar, R.M., Anoop, M., Harish, T., Rohit Prasad, A.K., Ramakrishnan, A.G.: DCT Based pitch modification, in Proc. SPCOM 6th Biennial Conf., pp. 114-117 (2001)
- Stylianou, Y.: Applying the harmonic plus noise model in concatenative speech synthesis, in Proc. IEEE Trans. Speech Audio Process., vol. 9, no. 1, pp. 21-29 (2001)
- Suni, A., Aalto, D., Raitio, T., Alku, P. and Vainio, M.: Wavelets for intonation modeling in HMM speech synthesis, in Proc. 8th ISCA Speech Synthesis Workshop, 2013. pp. 285-290 (2013)
- Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Speaker adaptation for HMM-based speech synthesis system using MLLR, Proceeding of the 3rd ESCA/COCOSDA workshop on Speech Synthesis, pp. 273-276 (1998)
- Tanaka, K., Abe, M.: A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F_0 , In Proc. ICASSP '97, Vol. 2, pp. 951-954 (1997)
- Tao, J.: F_0 prediction model of speech synthesis based on template and statistical method, in Lecture Notes of Artificial Intelligence, vol. 3206, pp. 497-504 (2004)
- Tokuda, K.: Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, Proc. of ICASSP, pp. 229-232 (1999)
- Tokuda, K., Zen, H., Black, A. W.: An HMM-based speech synthesis system applied to English, Proc. of IEEE Workshop on Speech Synthesis, pp. 227 - 230 (2002)
- Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K., Speech Synthesis Based on Hidden Markov Models, in Proceedings of the IEEE , vol.101, no.5, pp.1234-1252 (2013)
- Tóth, B., Németh, G.: Hidden Markov Model Based Speech Synthesis System in Hungarian, Infocommunications Journal, Volume LXIII, 2008/7, pp. 30-34 (2008)
- Tóth, B., Németh, G.: Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis, Acta Cybernetica, 19(4), pp. 715-731 (2010)
- Tóth, B., Nagy, P., Németh, G.: Towards Modeling Interrogative Sentences in HMM-based Speech Synthesis, the Phonetician, 109-101, 2014-I-II, pp. 24-42 (2014)
- Varga, L.: A hanglejtés [The Intonation], In Kiefer F. Structural Hungarian Grammar 2. Akadémiai Kiadó, pp. 468-546 (1994) (in Hungarian)
- Wang, Y., Jia, J., Cai, L.: Analysis of Chinese interrogative intonation and its synthesis in HMM-based synthesis system, International Conference on Internet Computing & Information Services (ICICIS), pp. 343-346 (2011)
- Wu, C. H. and Chen, J. H.: Automatic generation of synthesis units and prosodic information for chinese concatenative synthesis, Speech Communication, vol. 35, pp. 219-237 (2001)
- Xie, H., Andrae, P., Zhang, M., Warren, P.: Detecting stress in spoken English using Decision Trees and Support Vector Machines, in Proc. of the ACSW Frontiers 2004, Vol. 32, pp. 145-150 (2004)
- Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-based Speech Synthesis, IEICE Trans. Information and Systems, E88-D, no.3, pp.503-509 (2005)
- Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K.: The HTS-2008 system: Yet another evaluation of the speaker adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge, In Proc. Blizzard Challenge 2008 (2008)
- Yoshimura, T., Tokuda, K., Masuko T., Kobayashi T., Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, in Proc. Eurospeech, pp. 2347-2350 (1999)
- Yu, Y., Li, D., Wu, X., Prosodic modeling with rich syntactic context in HMM-based Mandarin speech synthesis, 2013 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), pp.132-136 (2013)
- Yuan, J., Shih, C., Kochanski, Greg P.: Comparison of declarative and interrogative intonation in Chinese, In Speech Prosody 2002, pp. 711-714 (2002)

- Zen, H., Tokuda, K., Masuko T., Kobayashi, T., Kitamura, T.: A hidden semi-Markov model-based speech synthesis system, *IEICE Trans. Inf. & Syst.*, E90-D(5), pp. 825-834 (2007)
- Zen, H., Tokuda, K., Black, A. W.: Statistical parametric speech synthesis, *Speech Communication* vol. 51, pp. 1039-1064 (2009)
- Zen, H., Senior, A., Schuster, M., Statistical Parametric Speech Synthesis using Deep Neural Networks, *International Conference on Acoustics, Speech and Signal Processing*, pp. 7962-7966 (2013)
- Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S.: Toward emotional speech synthesis: a rule based approach, In *Proc. 5th ISCA Speech Synthesis Workshop*, pp. 219-220 (2004)