INFORMATICA, 2016, Vol. 27, No. 3, 673–688

DOI: http://dx.doi.org/10.15388/Informatica.2016.105

Corpus-Based Hidden Markov Modelling of the Fundamental Frequency of Lithuanian

Airenas VAIČIŪNAS*, Gailius RAŠKINIS, Asta KAZLAUSKIENĖ
Vytautas Magnus University, K. Donelaičio g. 58, LT-44248, Kaunas, Lithuania
e-mail: airenass@gmail.com, g.raskinis@if.vdu.lt, a.kazlauskiene@hmf.vdu.lt

Received: March 2015; accepted: November 2015

Abstract. This paper presents the corpus-driven approach in building the computational model of fundamental frequency, or F_0, for Lithuanian language. The model was obtained by training the HMM-based speech synthesis system HTS on six hours of speech coming from multiple speakers. Several gender specific models, using different parameters and different contextual factors, were investigated. The models were evaluated by synthesizing F_0 contours and by comparing them to the original F_0 contours using criteria of root mean square error (RMSE) and voicing classification error. The HMM-based models showed an improvement of the RMSE over the mean-based model that predicted F_0 of the vowel on the basis of its average normalized pitch.

Key words: HMM-based synthesis, statistical parametric speech synthesis, F_0 modelling, intonation modelling, text-to-speech systems, HTS, Lithuanian language.

1. Introduction

An intonation expresses linguistic information about word meanings, indicates emphasized words of an utterance and helps to distinguish affirmative and interrogative statements. It also helps expressing paralinguistic information like emotions, identifying speaker and so on. A synthesis of the intonation plays an important role in text-to-speech (TTS) systems. A goal of such synthesis process is to predict the intonation for a text. The model of the intonation takes text labels and other linguistic and non-linguistic features as an input and produces the fundamental frequency (F_0) contour as an output. The model of the intonation $X()$ in a formal form can be expressed as:

$$X\big(f(1), f(2), \ldots, f(N)\big) \to F_0(1), F_0(2), \ldots, F_0(M), \tag{1}$$

where $f(1), f(2), \ldots, f(N)$ are vectors of contextual features extracted from a text each describing some chosen modelling unit, and $F_0(1), F_0(2), \ldots, F_0(M)$ represent the discrete time sequence of F_0 values henceforth referred to as F_0 contour generated for that text. The F_0 can take continuous value in a voiced region of a speech and is undefined in an unvoiced region of a speech: $F_0(x) \in \{NULL\} \cup (-\infty, \infty)$.

*Corresponding author.

1.1. *Intonation Synthesis Models*

There are many models and methods proposed and used to model the intonation. The first distinction is among the methods that split F_0 prediction process into a few steps and those that try to predict F_0 directly from the contextual features. The intonation modelling methods of the first type extract intonation model parameters P in the first step

$$X\big(f(1), f(2), \ldots, f(N)\big) \to P(1), P(2), \ldots, P(L)$$

and build intonation model $Y()$ that predicts F_0 contour in the second step

$$Y\big(P(1), P(2), \ldots, P(L)\big) \to F_0(1), F_0(2), \ldots, F_0(M).$$

For instance, a popular **Fujisaki** model (Fujisaki and Hirose, 1984) simulates a human intonation production mechanism. F_0 contour is defined as a linear superposition of a phrase intonation and accent intonation components:

$$\ln F_0(t) = \ln F_{0\min} + \sum_{i=1}^{I} A_{\mathrm{p}i} G_{\mathrm{p}i}(t - T_{0i}) + \sum_{j=1}^{J} A_{\mathrm{a}j}\big[G_{\mathrm{a}j}(t - T_{1j}) - G_{\mathrm{a}j}(t - T_{2j})\big],$$

where the second term of the expression is a slow decaying phrase intonation component and the last term of the expression is a fast decaying accent intonation. For a detailed explanation of the above expression see Fujisaki and Hirose (1984). Any classification method can be used to predict phrase and accent commands from a given text in the first step. Thereafter, the F_0 contour is calculated based on the above expression in the second step. The Fujisaki model was designed for Japanese but was successfully applied to many languages. The main advantage of this model is that it has constraints on the generated F_0 contour. It always generates a "naturally" looking F_0 contour.

A **Tilt** intonation model (Taylor, 1998) is modelling intonation at a phonetic level. It assumes that an intonation contour consists of a sequence of parametrized intonation events. Two types of events are defined: pitch accents and boundary tones. Each event is assigned a set of continuous parameter values a) amplitude ($A = |A_{\mathrm{rise}}| + |A_{\mathrm{fall}}|$) and b) duration ($D = |D_{\mathrm{rise}}| + |D_{\mathrm{fall}}|$). The F_0 contour in the region of the event is calculated using the following expression:

$$F_0(t) = \begin{cases} A_{\mathrm{abs}} + A - 2A(t/D)^2, & 0 < t \leqslant D/2, \\ A_{\mathrm{abs}} + A - 2A(1 - t/D)^2, & D/2 < t < D, \end{cases}$$

where A_{abs} is the absolute F_0 value at the start of the event and corresponds to the end value of the previous event or connection. The final F_0 contour is calculated by linearly interpolating F_0 in between the boundaries of the adjacent Tilt events. Tilt events and parameters of these events are predicted from contextual features. For example, Dusterhoff *et al.* (1999) successfully used decision trees for this task. The Tilt model has more degrees

of freedom than the Fujisaki model. Certain F_0 phenomena such as continuous rising of the intonation are hard to approximate with the Fujisaki model but possible with the Tilt model (Taylor, 2000).

A **Tones and Break Indices** (ToBI) system is an international standard for labelling intonation databases. It describes how speech corpora should be labelled. Black and Hunt (1996) proposed a method to generate F_0 contour by means of linear regression given the ToBI labelled speech corpus. The method predicts three F_0 values per syllable for three time positions within a syllable: at the start, in the middle and at the end of it. The prediction for each time position was made using separate linear regression models:

$$F_{0\,\text{point}} = I + w_1 f_1 + w_2 f_2 + \cdots + w_n f_n,$$

where f_i are features and I, w_i are model parameters estimated from the data by linear regression. The final F_0 contour is obtained by smoothing F_0 values at these three positions of every syllable. The main inconvenience of the ToBI model, when it is used for the synthesis of the F_0 contour, is related to the difficulty of obtaining ToBI labelled data. It is very hard to automatize ToBI label assignment (Taylor, 2009).

Unit concatenation TTS synthesis dominates today, but there is an increasing competition from statistical data driven systems, especially **HMM**-based systems. The existing open-source tools such as HTK (Young *et al.*, 2006) and HTS (Tokuda *et al.*, 2002b; Zen *et al.*, 2007a) offer a good starting platform for research and experiments on such systems. Though English and Japanese (HTS is created and supported in Japan) attracted most attention, other languages that are closer to Lithuanian were also investigated. HMM-based TTS systems are reported for highly inflected languages like Polish (Kuczmarski, 2010), Czech (Hanzliček, 2010), Croatian (Martincic-Ipsic and Ipsic, 2006) and others. The base HMM is described by the parameter set $\lambda = \{\pi, A, B\}$, where $\pi = \{\pi_i\}$ represents initial state probabilities, $A = \{a_{ij}\}$ – state transition probabilities and $B = \{b_i(o)\}$ – state output probabilities. Each chosen modelling unit (F_0 modelling is usually based of phonemes or syllables) is represented by a HMM having several states. The model is trained by trying to find the best parameter set λ:

$$\hat{\lambda} = \arg\max_{\lambda} p(O|W, \lambda);$$

$$p(O|W, \lambda) = \sum_{\forall q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t),$$

where $O = \{o_0, \ldots, o_T\}$ represents F_0 observations of the training data. Each observation can be a single F_0 value or can represent a vector including F_0 and its derivatives. W represents transcriptions (or contextual features) and $Q = \{q_0, \ldots, q_N\}$ represents an HMM state sequence. Then the output o is synthesized by:

$$\hat{o} = \arg\max_{o} p(o|W, \hat{\lambda}). \tag{2}$$

To model the discontinuous F_0 function, a multi space probability distribution MSD–HMM variation is mostly used (Tokuda *et al.*, 2002a). In the simplest case the output probability of an HMM state is defined by:

$$b_i(o) = w_i + (1 - w_i)\mathcal{N}(o; \mu_i, \sigma_i),$$

where w_i is a probability of the state i to produce unvoiced F_0, and $\mathcal{N}(o; \mu_i, \sigma_i)$ is a Gaussian distribution with mean μ_i and deviation σ_i to produce a voiced F_0 output o for the state i. In addition to F_0, delta of F_0 and acceleration of F_0 are used to model dynamics of the contour:

$$\delta F_0(t) \approx -\frac{1}{2}F_0(t-1) + \frac{1}{2}F_0(t+1),$$
$$\delta^2 F_0(t) \approx \frac{1}{4}F_0(t-1) - \frac{1}{2}F_0(t) + \frac{1}{4}F_0(t+1).$$

These three parameters are modelled in separate streams using the above defined MSD–HMM. A formula for output probability using streams is:

$$b_i(o) = \prod_{s=1}^{S} \left\{ b_i^{(s)}\left(o_i^{(s)}\right) \right\}^{w_s},$$

where S is the number of streams, $b^{(s)}(o^{(s)})$ is a state output probability of the stream s, and w_s is a stream weight, which defaults to one. There are also other methods proposed to model discontinuity of F_0, for example, by considering it to be continuous everywhere (Yu and Young, 2011).

Spectrum, intonation and duration are modelled and trained simultaneously in the HMM-based speech synthesis systems. Typically a semi-Markov structure is used by adding the duration model. Each HMM has the explicit state duration model instead of state transition probabilities. The duration is approximated by a Gaussian distribution and incorporated into training and synthesis parts (Zen *et al.*, 2007b). For other modifications of conventional HMM that are used in TTS systems see Tokuda *et al.* (2013).

A well-known problem is that an output generated by HMM systems suffers from over-smoothing. Toda and Tokuda (2007) introduced a method that uses a speech parameter generation considering Global Variance (GV). It adds a penalty to the maximized expression (Eq. (2)) if a probability density of the output differs from the global one estimated on all training data:

$$\hat{o} = \arg\max_{o} \left(p(o|W, \hat{\lambda})^{\omega} p(v(o)|\hat{\lambda}_{\mathrm{v}}) \right),$$

where $p(v(o)|\hat{\lambda}_{\mathrm{v}})$ is modelled by a single Gaussian distribution. If only F_0 contour is used, then $p(v(o)|\hat{\lambda}_{\mathrm{v}}) = \mathcal{N}(v(o); \mu_{\mathrm{v}}, \sigma_{\mathrm{v}})$. Where $v(o)$ is variation of synthesized contour, and $\mu_{\mathrm{v}}, \sigma_{\mathrm{v}}$ are mean and deviation of the contour estimated on the training set, ω is a weight to control the balance of two probabilities.

In recent days there were several different attempts to model F_0 by using HMMs. For example, Hirose *et al.* (2014) trained HMM system using an output of the Fujisaki model instead of real estimated F_0 values. Authors also incorporated F_0 residuals (fragments of F_0 contour not represented by the Fujisaki model) into training and synthesis steps. Suni *et al.* (2013) used wavelets to decompose F_0 contour into five temporal scales ranging from microprosody to the utterance level. Each F_0 level was trained individually and their superposition was used during the synthesis stage.

Further information about other intonation modelling methods can be found in detailed observations by Taylor (2009) and Santen *et al.* (2008).

1.2. *Contextual Factors for Intonation Modelling*

A set of linguistic and non-linguistic factors named contextual factors is defined to complement the description of a modelling unit. The factors are used to extract contextual features (denoted as $f(i)$ in the Eq. (1)). Contextual factors are useful for an intonation model because the same phone uttered in different context may result in different F_0 trajectories. The modelling unit often is chosen to be a syllable (Dusterhoff *et al.*, 1999) or a phoneme (Tokuda *et al.*, 2002b). Contextual factors that describe the context of the modelling unit at the level of phonemes, syllables, words, phrases and utterances are used for English (Tokuda *et al.*, 2002b). Features for some of these factors are easily estimated like the number of phonemes in a syllable or the position of the phoneme in a syllable, word, phrase and so on. Estimation of some other features, such as ToBI endtone label of an utterance or stress of a word, requires additional knowledge or robust estimation methods. Krstulovic *et al.* (2007) found that the performance of intonation models based on a set of objective factors is almost the same as the performance of the models based on complex factors like ToBI endtones. Cernak *et al.* (2013) investigated the relative importance of contextual factors for English and found that the syllable context is the most important and the word context is less important. The factors that influence the intonation are language dependent. Other factors related to Japanese language are described by Yoshimura *et al.* (1999), to Czech – by Hanzliček (2010).

1.3. *Research on Intonation Modelling of Lithuanian Language*

The problem of creating computational models of Lithuanian intonation has already been addressed by some researchers. Paulikas and Navakauskas (2005) present the third order polynomial accent models of F_0 and incorporate them into the homograph restoration process of Lithuanian words. Leonavičius (2006) models pitch variation of melismas using artificial neural networks. The authors of the Lithuanian TTS synthesis system (Kasparaitis, 2001; Kasparaitis and Anbinderis, 2014) probably use some rule-based F_0 prediction techniques, though the details of their F_0 modelling are not revealed. Our approach to F_0 modelling is different from the research above in two aspects. First, our research is not limited to some particular context (e.g. homographs, melismas) of the spoken language and has the ultimate goal of predicting the F_0 contour in as many contexts as possible. Second, our research is driven by the large corpus of spoken utterances that intend to approximate the actual usage of the language.

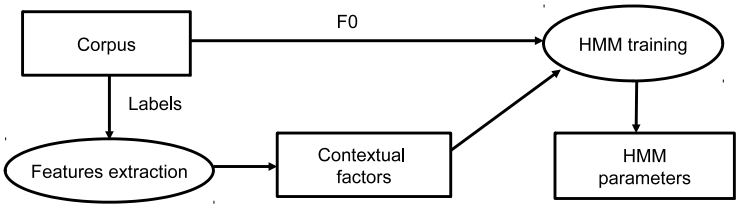

Fig. 1. HMM training process.

2. Intonation Modelling of Lithuanian Language Using HMMs

In this section we present HMM-based intonation models of Lithuanian language. The models were built using a simplified version of HTS 2.2 toolkit (see online HTS Web Site). We followed the default sequence of steps described in *CMU-ARCTIC-SLT* demo to train the intonation models. A high level abstraction of the training process is illustrated by Fig. 1. Only F_0 and derivatives of F_0 were used to prepare training parameters. Spectrum parameters were not used in the experiments. Phoneme was chosen to be the modelling unit. SAMPA-LT (Raškinis *et al.*, 2003) phoneme inventory was taken for the basic phoneme set. This set distinguishes between stressed/unstressed, short/long vowels, and palatalized/non-palatalized consonants. Palatalization of consonants was ignored as it can be captured by using wider phoneme context. This simplification resulted in a reduced inventory of about 100 phonemes with respect to the original SAMPA-LT set. A short and a long pause symbols were also included into the set of modelling units. Context-dependent phonemes were modelled using MSD-HMM states with no skips. F_0, delta and acceleration parameters were modelled in separate streams.

The list of contextual factors used in our experiments is presented in Table 1. It was derived from the list of factors that were used for English (Tokuda *et al.*, 2002b). Almost all of these factors are objective and the contextual features can be obtained automatically. A few subjective, manually labelled, semantic factors such as rheme, theme and logical stress were added to this set in a hope that they will increase the predictive accuracy of the model. The total number of contextual factors was 49.

A rich list of contextual factors results in that almost every feature vector becomes unique and parameters of HMM states cannot be accurately estimated. State tying technique is applied to overcome the data sparseness problem. Similar states are clustered together using hierarchical decision tree. A minimum description length criterion is used to build the tree (Shinoda and Watanabe, 2000). The state tying algorithm uses a set of factor value tests to partition training set into two parts on each step. Factors with phoneme labels as their values were subject to the membership tests to some category (set of phoneme symbols). Such tests are language specific so considerable amount of time was spent on defining an extensive set of Lithuanian-specific phonetic categories. Numeric valued factors were subject to the tests of type "=" and "$\leqslant$". "Part-of-speech" factor was a subject to the "=" type of tests.

Table 1
Contextual factors used to describe the context of the modelling unit (phoneme).

Relates to	Features
Phoneme	label of the current phoneme, labels of two preceding and two succeeding phonemes, forward, backward positions of the current phoneme in a syllable
Previous, current and next syllable	stress presence, stress type, number of phonemes
Current syllable	forward, backward positions of the syllable in a word and a phrase, vowel of the syllable, number of syllables that are stressed in a phrase before and after the current syllable, number of syllables from previous stressed syllable, number of syllables to next stressed syllable
Previous, current and next word	number of syllables, part of speech tag, logical stress
Current word	forward, backward positions of the word in a phrase, distance in words to comma, dot, question mark or exclamation mark
Previous, current and next phrase	number of syllables, words
Current phrase	forward, backward positions of the phrase in an utterance rheme, theme
Utterance	number of syllables, words, phrases

Table 2
The speech corpus.

Gender	Speakers	Train (min)	Test (min)
Male	10	126.78	19.25
Female	9	181.23	29.99

3. Model Assessment

3.1. *Speech Corpus*

Lithuanian speech corpus was used in our investigations. It contained utterances of read speech of 10 males and 9 females, nearly 6 hours of speech in total. The speech corpus was annotated at the level of phonemes, syllables, words, and phrases. It also included some semantic and morphological markup: boundaries of themes, rhemes and logical stress, and part of speech tags.

Speaker independent, but gender dependent models were built and investigated in our experiments. Table 2 describes training and testing parts of the corpus.

3.2. F_0 Contour Extraction and Normalization

F_0 contour was extracted from audio signals using the procedure described in Raškinis and Kazlauskienė (2013). F_0 values were calculated using 5 ms frame shift. The F_0 contour was normalized for all speakers:

$$\ln F_{0\,\mathrm{norm}}(t) = \ln F_{0\,\mathrm{speaker}}(t) - \overline{\ln F_{0\,\mathrm{speaker}}} + \overline{\ln F_{0\,\mathrm{all}}},$$

where $\overline{\ln F_{0\,\mathrm{speaker}}}$ and $\overline{\ln F_{0\,\mathrm{all}}}$ are mean of logarithmic F_0 values for a speaker and for the entire corpus respectively. The initial size of the speech corpus was greater than that indicated by Table 2. Some utterances were excluded from training and testing corpora because the extracted F_0 contour might be erroneous. Criteria for automatic rejection of utterances were based on checking F_0 contour for a) an octave jump within a short period of time and b) a range of 4 octaves within one utterance.

3.3. *Evaluation Criteria*

The F_0 contour was synthesized on the basis of text transcriptions of utterances in the test corpus. The synthesis process used the knowledge of phoneme boundaries. Thus, the synthesized/predicted F_0 contour was of the same length as the reference F_0 contour of the test corpus and a frame by frame comparison could be performed. For the evaluation a root mean square error (RMSE) and a voicing classification error (VCE) criteria were used. The RMSE is calculated taking the quadratic difference for each frame:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t\in V}\left(\ln F_0(t) - \ln \widehat{F_0}(t)\right)^2}{N_V}},$$

where N_V is the total number of frames where both reference F_0 values and predicted $\widehat{F_0}$ estimates are voiced.

The VCE is a percent of voiced and unvoiced frames for which the wrong voicing decision was taken:

$$\mathrm{VCE} = 100\frac{N_{\mathrm{wrong}}}{N_{\mathrm{all}}},$$

where N_{wrong} is a number of wrong voicing decisions and N_{all} is a total number of frames within the testing material.

3.4. *Baseline Mean-Based Model*

A simple static mean-based model was used as a starting point for an evaluation of the models. A single mean value of the F_0 was extracted for each voiced phoneme. These values were calculated on a multi-speaker normalized corpus. F_0 synthesis consisted of mapping every voiced phoneme of the test corpus into its constant mean value. A linear

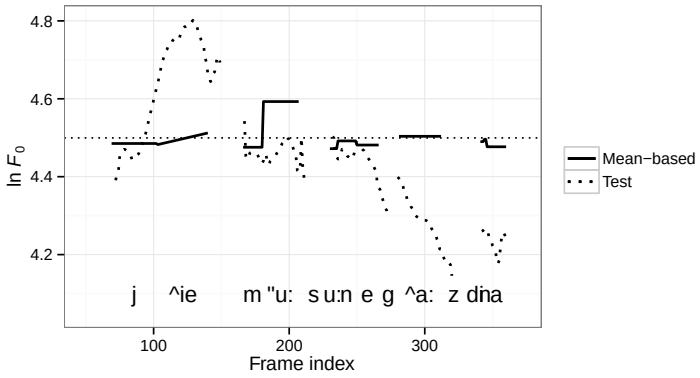

Fig. 2. Sample of contour generated by the mean-based model.

Table 3
Model evaluation results.

Model type	Male		Female	
Features & contextual factors	RMSE	VCE, %	RMSE	VCE, %
1. Baseline mean-based[a]	0.21465	12.23	0.17918	13.19
2. F_0 & full set[b]	0.22493	13.66	0.18625	13.06
3. F_0+delta & full set	0.20535	11.34	0.17306	11.75
4. F_0+delta+acc & full set	0.19952	12.09	0.16873	10.75
5. F_0+delta+acc & full set & GV[c]	0.22373	12.05	0.19538	10.75
6. F_0+delta+acc & without subjective[d]	0.19921	12.02	0.16855	10.73
7. F_0+delta+acc & only phonemes	0.20869	11.92	0.17366	10.66

[a] Static non HMM-based model.
[b] See Table 1 for full set of contextual factors.
[c] Using GV method.
[d] Full set excluding subjective factors: rheme, theme and logical stress.

interpolation was used to calculate F_0 values for each diphthong. A sample of the contour generated by this method is provided in Fig. 2. The RMSE and VCE results for this type of model are provided in Table 3 (row 1).

3.5. *Default HMM-Based Model*

Gender dependent models were built using the training procedure, described in Section 2. Parameters included F_0, delta and acceleration of F_0. A model topology was based on 5 generating left-to-right states per phoneme. State skip was not allowed. 1 mixture per state was used. A full set of contextual factors was included (see Table 1). Global Variance method was not used in a synthesis stage. The results (Table 3, row 4) showed an improvement comparing to the baseline model.

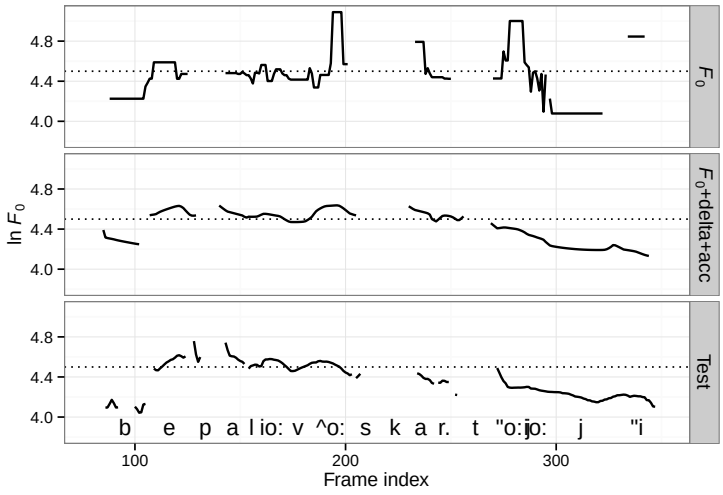

Fig. 3. Sample of contours generated by HMM with and without dynamic features.

3.6. *Dynamic Features*

To experimentally prove the importance of dynamic features we calculated the RMSE for models without delta and acceleration of F_0. The results (Table 3, rows: 2, 3) showed that the use of dynamic features is crucial for this type of models. Fig. 3 shows samples of F_0 contours that were generated with and without the dynamic features.

If dynamic features are not used, then F_0 contour just follows the mean F_0 value of individual HMM states. When F_0 derivative was added, it greatly improved the RMSE value. The inclusion of the F_0 acceleration (Table 3, row 4) added another improvement.

3.7. *Global Variance Method*

The speech synthesis can be audibly improved by using the GV method. The most significant improvement results from applying GV to the spectral synthesis and to a lesser extent from applying it to the F_0 synthesis (Toda and Tokuda, 2007). We evaluated a default HMM-based model by applying the GV method for the F_0 synthesis. The estimated RMSE values of this experiment are provided in Table 3 (row 5).

Distributions of the reference F_0 values and the synthesized F_0 values are shown in Fig. 4. As it could be expected, the distribution of synthesized F_0 values using the GV method is closer to the reference distribution. Surprisingly, the RMSE criterion showed an opposite result. By exploring the synthesized F_0 contours in detail we found that GV has the tendency to "cheat". Sometimes it generates a few exceptional F_0 values in order to increase a variance of an utterance (see Fig. 5). This phenomenon can be explained by the hypothesis that "flat" probability density functions were learnt (or were not) by some HMM states. Flat distributions do not penalize synthesized F_0 values that are far away from the mean F_0 value of a state.

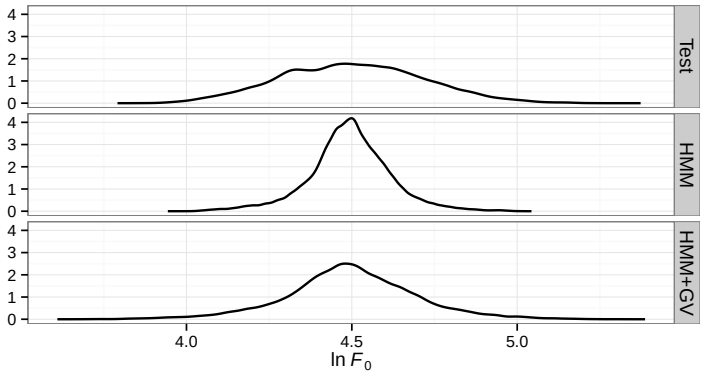

Fig. 4. F_0 densities of test and predicted contours (male speakers).

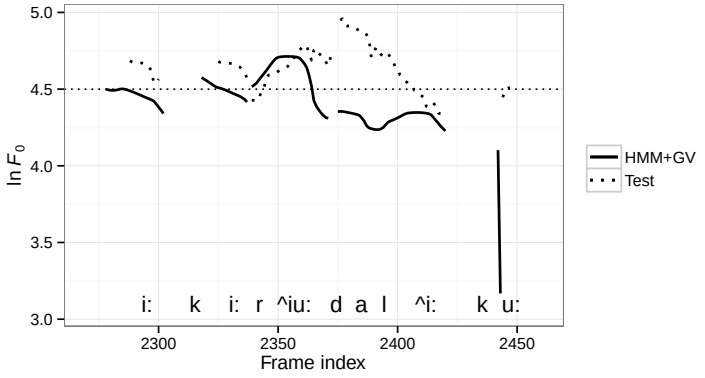

Fig. 5. Sample of contour generated by HMM with GV method.

3.8. *Impact of the Number of States per Phoneme*

By default, HTS suggests using HMM topology with 5 generating states without skips for each phoneme. HMMs with more states per phoneme may allow for a more fine-grained F_0 modelling. However, because of no-skips HMM topology, short phonemes may not be modelled accurately. The minimum phoneme duration that can be accurately modelled with 5 HMM states is 25 ms (F_0 was calculated with 5 ms frame shift) and 50 ms with 10 HMM states. There are 9% and 35% of phonemes that have the duration $\leqslant$ 25 ms and $\leqslant$ 50 ms respectively in our corpus. The percentage of frames covered by these phonemes is a bit lower – 2.7% and 18% respectively.

The impact of the number of states per phoneme was investigated by building F_0 models on the basis of different HMM topologies. Fig. 6 shows the results of this investigation. The results confirm the fact that more detailed F_0 contour is generated if more states are used. But the VCE increases for models with more states.

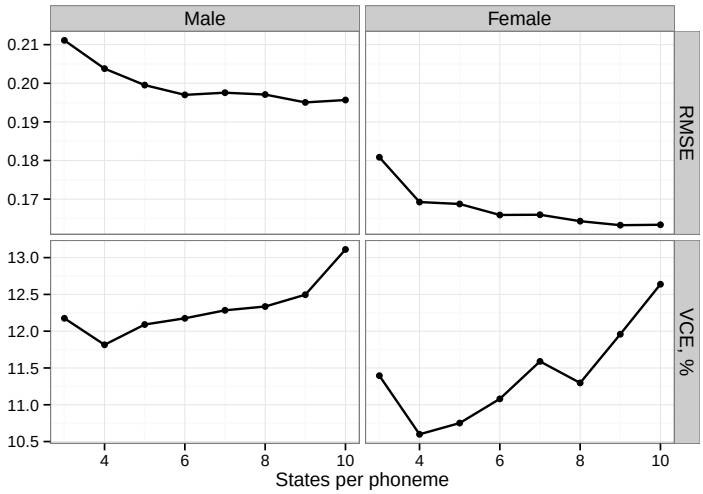

Fig. 6. RMSE and VCE estimates for models with different number of HMM states per phoneme.

3.9. *Impact of Some Contextual Factors*

We conducted an experiment that tried to estimate the impact of the subjective contextual factors: theme, rheme and logical stress on model accuracy. We constructed a model that used the complete set of factors (see Table 1) except for theme, rheme and logical stress. The RMSE of this model (Table 3, row 6) was a bit lower in comparison to the default HMM-based model.

The analysis of the decision tree generated in order to tie HMM states of the default HMM-based model revealed that node splitting tests of the type "Is theme" and "Is rheme" were among top 10 most frequent tests. This suggests that the dependency between theme, rheme and F_0 contour really exists. At the same time, it seems that if these factors are missing, the other objective factors (Table 1) can successfully compensate for them.

A model with a small set of contextual factors (current and two preceding and two succeeding phonemes) showed a big increase in the RMSE value (Table 3, row 7). This confirms that the selection of right contextual factors is important and can improve the quality of the synthesized F_0 contour.

4. Conclusions

A few HMM-based F_0 prediction models were built and experimentally investigated for Lithuanian language. The default HMM-based model achieved lower RMSE estimate in comparison to the static baseline model. The absolute improvement was small, but the difference was statistically significant and confirmed by paired t-test with confidence interval 0.95. We were unable to perform cross-validation tests due to the high amount of computations involved in such tests.

The experiments with different model types showed the usefulness of the delta and acceleration features. The increase of states per phoneme from the default number (5 generating states) resulted in a reduced RMSE value, but at the same time VCE increased. The GV method increased the RMSE and did not help to model the intonation.

We experimented with a few different sets of contextual factors. The set of contextual factors that was good for English appeared to be useful for Lithuanian as well. However, our assumption that Lithuanian intonation model can be improved by the subjective factors: theme, rheme and logical stress, was not confirmed. A detailed investigation of additional Lithuanian-specific contextual factors that could improve over the present contextual factor set is one of the directions of our future research.

Further investigation of different HMM topologies for different phonemes may be promising. For instance, short and long Lithuanian vowels may be modelled by HMMs with different number of states. Allowing state skip for certain phonemes may also result in a more accurate F_0 models. We expect that F_0 models can be also improved by providing more training data. Finally, the subjective listening tests of synthesized F_0 contours can be performed. If correlation of listener ratings and RMSE values is established, it may help to better interpret the achieved RMSE results.

Acknowledgement. This research was funded by a grant (No. LIT-5-4) from the Research Council of Lithuania.

References

Black, A.W., Hunt, A.J. (1996). Generating F0 contours from ToBI labels using linear regression. In: *Proceedings of ICSLP*, Vol. 3. IEEE, pp. 1385–1388.

Cernak, M., Motlicek, P., Garner, P.N. (2013). On the (UN)importance of the contextual factors in HMM-based speech synthesis and coding. In: *Proceedings of ICASSP*, pp. 8140–8143.

Dusterhoff, K.E., Black, A.W., Taylor, P. (1999). Using decision trees within the Tilt intonation model to predict F0 contours. In: *Proceedings of EUROSPEECH*, pp. 1627–1630.

Fujisaki, H., Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4), 233–242.

Hanzliček, Z. (2010). Czech HMM-based speech synthesis. In: Sojka, P., Horak, A., Kopeček, I., Pala, K. (Eds.), *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, Vol. 6231. Springer, Berlin, pp. 291–298.

Hirose, K., Hashimoto, H., Hyakutake, K., Saito, D., Minematsu, N. (2014). Use of fundamental frequencies shaped by generation process model for hmm-based speech synthesis. In: *Proceedings of International Conference on Signal Processing (ICSP)*, pp. 555–560.

Hts web site [online]. (2014) [cited 2014-12-01]. Available from: http://hts.sp.nitech.ac.jp/.

Kasparaitis, P. (2001). *Lithuanian Text-to-Speech Synthesis*. PhD thesis, Vilnius University.

Kasparaitis, P., Anbinderis, T. (2014). Building text corpus for unit selection synthesis. *Informatica*, 25(4), 551–562.

Krstulovic, S., Hunecke, A., Schröder, M. (2007). An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements. In: *Proceedings of Interspeech*, pp. 1897–1900.

Kuczmarski, T. (2010). HMM – based speech synthesis applied to Polish. In: Demenko, G., Wagner, A. (Eds.), *Speech and Language Technology*, Vols. 12/13. Polish Phonetic Association, pp. 221–228.

Leonavičius, R. (2006). *Melisma synthesis using artificial neural networks*. PhD thesis, Vilnius Gediminas Technical University.

Martincic-Ipsic, S., Ipsic, I. (2006). Croatian HMM based speech synthesis. In: *Proceedings of International Conference on Information Technology Interfaces*, pp. 251–256.

Paulikas, Š., Navakauskas, D. (2005). Restoration of voiced speech signals preserving prosodic features. *Speech Communication*, 47(4), 457–468.

Raškinis, A., Raškinis, G., Kazlauskienė, A. (2003). SAMPA (Speech Assessment Methods Phonetic Alphabet) for encoding transcriptions of Lithuanian speech corpora. *Information Technology and Control*, 4(29), 52–55.

Raškinis, G., Kazlauskienė, A. (2013). From speech corpus to intonation corpus: clustering phrase pitch contours of Lithuanian. In: *Proceedings of 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pp. 353–363.

Santen, J., Mishra, T., Klabbers, E. (2008). Prosodic processing. In: Benesty, J., Sondhi, M., Huang, Y. (Eds.), *Springer Handbook of Speech Processing*. Springer, Berlin, pp. 471–488.

Shinoda, K., Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *Acoustical Science and Technology*, 21(2), 79–86.

Suni, A., Aalto, D., Raitio, T., Alku, P., Vainio, M. (2013). Wavelets for intonation modeling in hmm speech synthesis. In: *Proceedings of 8th ISCA Workshop on Speech Synthesis*, pp. 285–290.

Taylor, P. (1998). The tilt intonation model. In: *Proceedings of ICSLP*, pp. 1383–1386.

Taylor, P. (2000). Analysis and Synthesis of Intonation using the Tilt model. *Journal of the Acoustical Society of America*, 107, 1697–1714.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge. Available from: http://books.google.lt/books?id=T0O-NHZx7kIC.

Toda, T., Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE – Transactions on Information and Systems*, 816–824.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T. (2002a). Multi-space probability distribution HMM. *IEICE – Transactions on Information and Systems*, 85(3), 455–464.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K. (2013). Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, 101(5), 1234–1252.

Tokuda, K., Zen, H., Black, A.W. (2002b). An HMM-based speech synthesis system applied to English. In: *Proceedings of 2002 IEEE Workshop on Speech Synthesis*. IEEE, pp. 227–230.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: *Proceedings of EUROSPEECH*, pp. 2347–2350.

Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge.

Yu, K., Young, S. (2011). Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5), 1071–1079.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K. (2007a). The HMM-based speech synthesis system (HTS) version 2.0. In: *Proceedings of ISCA Workshop on Speech Synthesis (SSW-6)*, pp. 294–299.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (2007b). A hidden semi-Markov model-based speech synthesis system. *IEICE – Transactions on Information and Systems*, 90(5), 825–834.

A. Vaičiūnas received PhD in the field of informatics in 2006. Since then he has worked as software engineer and researcher in various computational linguistics projects. His research interests are human language technologies.

G. Raškinis received his MSc degree in artificial intelligence and pattern recognition from the University of Pierre et Marie Curie in Paris in 1995. He received doctor's degree in the field of informatics (physical sciences) in 2000. Presently, he works at the Center of Computational Linguistics and teaches at the Department of Applied Informatics of Vytautas Magnus University. His research interests include application of machine learning techniques to human language processing.

A. Kazlauskienė received doctor's degree in the field of humanities (philology) in 1998. She teaches at the Department of Lithuanian Language of Vytautas Magnus University. Her research interests are phonology, phonotactics, accentuation, rhythm, applied linguistics.

Lietuvių kalbos pagrindinio tono kaitos prognozė, naudojant paslėptųjų Markovo modelių metodiką

Airenas VAIČIŪNAS, Gailius RAŠKINIS, Asta KAZLAUSKIENĖ

Šiame straipsnyje, remiantis 20 profesionalių diktorių 6 valandų skaitomos kalbos garsyno duomenimis, aprašomi bandymai automatizuotai sudaryti kompiuterinį lietuvių kalbos intonacijos modelį. Intonacijos modeliui kurti naudotas šnekos sintezės įrankis HTS, kuris remiasi Paslėptųjų Markovo Modelių (PMM) metodika. Tyrimuose atskirai modeliuoti vyrų ir moterų balsai, įvertinta įvairių HTS sistemos parametrų ir kontekstinių požymių įtaka prognozuojamos pagrindinio tono kreivės tikslumui. Gauti intonacijos modeliai vertinti lyginant jų prognozuotą pagrindinio tono kreivę su atskaitos kreivėmis. Vertinti naudoti vidutinės kvadratinės paklaidos ir balsingų / nebalsingų atkarpų atpažinimo klaidos kriterijai. Taikant PMM pagrindu veikiantį intonacijos modelį gaunamas mažesnis vidutinės kvadratinės paklaidos įvertis nei taikant paprastąjį modelį, kai kiekvienam balsiui atvaizduoti pasirenkama tam balsiui būdinga vidutinė pagrindinio tono reikšmė.