ESUKA – JEFUL 2015, 6–3: 145–168

EXPRESSION OF BASIC EMOTIONS IN ESTONIAN PARAMETRIC TEXT-TO-SPEECH SYNTHESIS

Kairi Tamuri and Meelis Mihkla
Institute of the Estonian Language

Abstract. The goal of this study was to conduct modelling experiments, the purpose of which was the expression of three basic emotions (joy, sadness and anger) in Estonian parametric text-to-speech synthesis on the basis of both a male and a female voice. For each emotion, three different test models were constructed and presented for evaluation to subjects in perception tests. The test models were based on the basic emotions' characteristic parameter values that had been determined on the basis of human speech. In synthetic speech, the test subjects most accurately recognized the emotion of sadness, and least accurately the emotion of joy. The results of the test showed that, in the case of the synthesized male voice, the model with enhanced parameter values performed best for all three emotions, whereas in the case of the synthetic female voice, different emotions called for different models: the model with decreased values was the most suitable one for the expression of joy, and the model with enhanced values was the most suitable for the expression of sadness and anger. Logistic regression was applied to the results of the perception tests in order to determine the significance and contribution of each acoustic parameter in the emotion models, and the possible need to adjust the values of the parameters.

Keywords: Estonian, emotions, speech synthesis, acoustic model, speech rate, intensity, fundamental frequency

DOI: http://dx.doi.org/10.12697/jeful.2015.6.3.06

1. Introduction

Modern text-to-speech synthesis is applied in various domains: in tools for people with hearing, reading, or speech disabilities, human-machine communication, multimedia products, computer games, etc. It is therefore increasingly important that synthetic speech sounds natural, as much like human speech as possible in all its aspects. One of the aspects of human speech is emotion – emotions are always present in human speech and it should therefore be present in synthetic speech as well.

In order to introduce emotions into synthetic speech, an acoustic model of emotional speech is necessary. The model provides the synthesizer with an appropriate combination of acoustic parameters, which is specific to each emotion. The application of acoustic models of emotions' prosodic parameters in parametric speech synthesis has been tested for many languages (Iriondo et al. 2004, Audibert et al. 2005). However, people speaking different languages and living in different cultural surroundings perceive and express emotions differently (Altrov and Pajupuu 2015, Paulmann and Uskul 2014, Altrov 2013). Therefore, the results obtained for other languages cannot be automatically adopted into Estonian. The Estonian speech synthesizer needs a model that is specific to Estonian. With this goal in mind, the Institute of the Estonian Language initiated, in 2008, a study of the acoustics of Estonian emotional speech. The aim was to identify the acoustic features that characterize three emotions – joy, sadness, and anger – and distinguish them from each other and from neutral speech, and that could be used to create an acoustic model of Estonian emotional speech for an Estonian speech synthesizer.

Joy, sadness, and anger were chosen for modelling because they belong to the category of basic emotions and can be easily described in terms of the variation of acoustic features (Ekman 1992, Scherer 2013). Also, the acoustic expression of these three emotions in Estonian has been previously studied. One of the authors of this paper, Kairi Tamuri, has analysed the acoustics of Estonian emotional speech read-aloud (see section 2), using the speech data in the Estonian Emotional Speech Corpus,[1] and determined the acoustic features and the values that characterize the three basic emotions in Estonian speech, and which distinguish them from one another and from neutral speech. She has analysed pauses (Tamuri 2010), formants and precision of articulation (Tamuri 2012a), speech rate (Tamuri and Mihkla 2012), intensity of speech (Tamuri 2012b) and fundamental frequency (Tamuri 2015) in emotional speech. It is important to underline that the object of study was not spontaneous or acted speech, but read-aloud speech. Similarly, the speech synthesizer that was used to test the modelling of emotions is a text-to-speech synthesizer, designed to read aloud written text.

The goal of the present study was to test whether and to what extent the characteristic acoustic parameter values of emotions that were identified on the basis of human speech are able to create emotions in Esto-

1 See http://peeter.eki.ee:5000/?lg=en

nian synthetic speech. Our aim was to create a simple acoustic model of emotional speech that is applicable in the parametric synthesis of Estonian speech, that is suitable both for a male and a female synthetic voice, and that uses four parameters (speech rate, loudness of voice, pitch level, and pitch range) – thereby helping the synthesizer express joy, sadness and anger. In the construction of the acoustic model, we took into account the results of the acoustic analysis of human speech, as well as the possibilities of the existing Estonian speech synthesizer. We used a statistical parametric HTS speech synthesizer based on hidden Markov models, which allows to directly control the above four acoustic parameters during the process of synthesis (Zen et al. 2007). For each emotion, three different test models were constructed and presented to subjects for evaluation in perception tests. Logistic regression was applied to the results of the perception tests in order to determine the significance and contribution of each acoustic parameter in the emotion models, and the possible need to adjust the values of the parameters.

2. The acoustic analysis of Estonian emotional speech

The creation of the acoustic model of Estonian emotional speech was based on the results of the acoustic study of Estonian human emotional speech (see below). Our goal was to determine for each emotion (joy, sadness, and anger) in which direction and to what degree their acoustic parameter values should be shifted in comparison with neutral speech and the other emotions.

2.1. The rate of Estonian emotional speech read-aloud

Tamuri and Mihkla (2012) analysed the speech rate of Estonian emotional speech and found that the rate is fastest in sentences expressing the emotion of anger, followed by sentences expressing the emotion of joy and neutral sentences; the speech rate is slowest in sentences expressing sadness. Speech rate differences were statistically significant between the pairs: anger *vs.* joy, anger *vs.* sadness, anger *vs.* neutral speech, and joy *vs.* sadness.

2.2. The intensity of Estonian emotional speech read-aloud

The results of Tamuri (2012b) on the intensity – i.e. the loudness – of Estonian emotional speech showed that in Estonian, the voice is loudest in neutral speech, followed by angry and happy speech, and the least loud in sad speech. The differences in intensity were statistically significant both between the pairs of emotions as well as in comparison with neutral speech.

2.3. The level and range of fundamental frequency in Estonian emotional speech read-aloud

According to the results of Tamuri (2015) on the fundamental frequency – i.e. the voice pitch – in Estonian emotional speech, the pitch is highest in sentences expressing the emotion of joy, followed by neutral and sad speech; the pitch is lowest in angry sentences. The most distinctive emotion was anger, whose difference from the other emotions and neutral speech was also statistically significant. The analysis of the range of fundamental frequency – i.e. of the variation of pitch (the difference between the maximal and the minimal value) – showed that in Estonian emotional speech, F0 range is largest in case of anger, followed by the emotion of joy and by neutral speech. The F0 range is narrowest in sad sentences. The differences in F0 range were statistically significant between the pairs: anger *vs.* sadness, anger *vs.* neutral speech, and joy *vs.* sadness.

3. Estonian speech synthesizers

An overall description of the existing Estonian speech synthesizers (Mbrola-et, UnitSelection-et, eSpeak-et and HTS-et) and their characteristics can be found in Mihkla et al. (2012) and Mihkla et al. (2013). These speech synthesizers are based on different methods and have different characteristics, each having its specific range of use. The most accurate segmental quality of speech is achieved by the synthesis method based on the selection of speech units (UnitSelection-et). The fluency and prosody of speech are most accurate in the synthesis system using hidden Markov models (HTS-et). The eSpeak system (eSpeak-et) has the most limited text processing possibilities, but is at the same time the most compact synthesis module. The unit selection system

offers limited possibilities in terms of the control and supervision of the synthesis process, but yields the richest variability of speech. All these synthesizers are constantly being further developed (except the Mbrola-et module, with respect to which, unfortunately, the development of the synthesis engine was terminated almost five years ago), and new synthetic voices are being generated and the existing ones improved.

For the purposes of the parametric synthesis of emotional speech, the unit selection-based speech synthesis (UnitSelection-et) is not suitable, because most parameters (including speech rate, F0, and loudness of speech) cannot be directly controlled under this method. Formant synthesis (eSpeak-et) in turn yields low segmental quality and a machine-like output speech, leading to a low recognisability of emotions.

We tested the parametric synthesis of emotions using both the method based on hidden Markov models (HTS-et) and the diphone synthesis (Mbrola-et). However, the Mbrola diphone synthesis does not permit a direct control of the F0 range with a special parameter; furthermore, the development of this synthesis engine has been terminated. Given that HTS-et is currently the most actively developed system and that its directly controllable parameters include the four parameters that were of interest to us (i.e. speech rate, speech intensity, F0, and F0 range), the HTS-et synthesizer was chosen for the parametric modelling of emotions. In addition to a parametric synthesis of emotional speech, this statistical parametric synthesis method is also suitable for a corpus-based approach; for instance, corpora of emotional speech are used to train speech models (Yamagishi et al. 2005, Lorenzo-Trueba et al. 2015). We used two synthetic voices trained on neutral speech, a male voice, Tõnu, and a female voice, Eva, as the basic voices of the Estonian HTS synthesis.

4. Parametric test models of emotional speech for text-to-speech synthesis

The parametric test models of emotional speech are based on the model of synthetic voices created with the HTS method and trained on corpora of neutral speech. The parameter values of the neutral synthetic speech are determined by the speech model of the statistical parametric synthesizer itself and are independent of the results on neutral speech obtained in the study of emotional speech.

The earlier acoustic analysis of Estonian human emotional speech was conducted on the basis of a female voice only. Therefore, the parameter values of the test models of the emotions could not be derived directly from the results of this analysis, given that our goal was to create models that are suitable both for a male and a female synthetic voice. Taking into account the results of the acoustic analysis of human emotional speech, as well as the parameter tuning possibilities of the particular speech synthesizer, we set out to identify, in a four-dimensional acoustic space, the values of the acoustic parameters (speech rate, pitch, intensity and pitch range) that would shift the synthesizer from the area of neutral speech to that of the characteristic values of the three basic emotions (joy, sadness, anger).

First, we constructed a parametric acoustic test model for each emotion. These models were then validated by a small group of experts (persons participating in the development of speech synthesis and the study of emotional speech). They evaluated the choice of the parameters and also suggested some changes in the parameters. On the basis of the experts' mean evaluations and their suggestions concerning the choice of parameters, we constructed the so-called reference models of the emotions (see Table 1, Model 2).

In addition to the reference models, two further models were created for each emotion (Model 1 and Model 3). The parameter values of Models 1 and 3 were derived from two-dimensional acoustic feature spaces (Figures 1 and 2). Figure 1 represents the parameter values of neutral speech and of the different models of the three emotions in the acoustic space of the speech rate and intensity. The speech rate and the intensity of neutral speech are represented by the relative values [1.0; 1.0]. The parameter values of the emotions are given in relation to the parameters of neutral speech: for all three emotions, the intensity of speech is lower than in neutral speech, and for sad speech, the speech rate is slower than in neutral speech, whereas for the other emotions it is faster. In this feature space, the parametric models of joy and anger are located relatively close to each other. Figure 2 represents the parameters of the models in the acoustic space of the F0 level and the F0 range. The parameters of neutral speech – [0.0; 0.0] – serve as the reference point for the other models, the F0 level and F0 range being either increased or decreased with respect to the model of neutral speech. For instance, the F0 levels of sadness are 3–5 semitones lower than those of the neutral model, whereas in the models of happy speech, the F0 is 1.5–3.5 semitones higher than in neutral speech (see Table 1). For the emotion of

anger, the F0 range is 2.1–3.1 semitones wider than for neutral speech, whereas in sad speech it is 0.9–1.4 semitones narrower than in neutral speech (see Table 1). In the acoustic space of the F0 level and F0 range, the models of all three emotions are relatively clearly separated from one another.

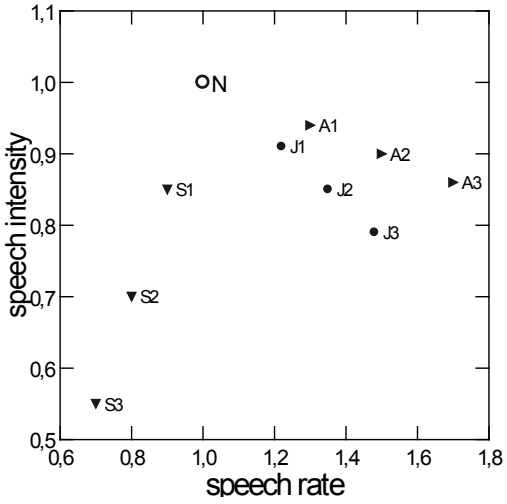

Figure 1. The relative parameter values of the models for neutral (N) and emotional speech (A – anger, J – joy, S – sadness) in the acoustic space of the speech rate and speech intensity.

The purpose of creating three different models for joy, sadness, and anger was to identify the acoustic limits in which a particular emotion model operates – i.e. to verify whether the initial model M2 had been determined with sufficient accuracy for the synthesizer to be able to synthesize joy, sadness, and anger in a recognizable manner, or whether it was necessary to further adjust the values of the parameters. Model 1 (M1) is a model with decreased values, where the values of M2 have been lowered by approximately 15% in the direction characterizing the emotion (towards neutrality). Model 3 (M3) is a model with increased values, in which the values of M2 have been raised by 15% in the emotion's characteristic direction (away from neutrality).

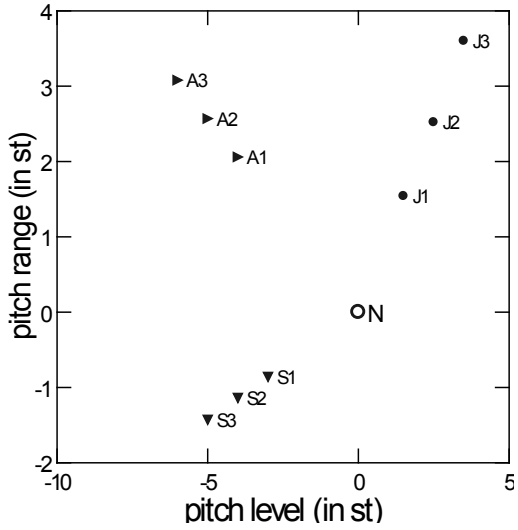

Figure 2. Parameter values (in semitones) of the models for neutral (N) and emotional speech (A – anger, J – joy, S – sadness) in the acoustic space of the pitch level and pitch range.

Table 1 presents the values of the parameters in relation to the default parameters of the neutral synthetic voice serving as the base model ([1.0, 1.0, 0.0, 0.0]). Speech rate and speech intensity, or loudness, are represented with the parameter value 1.0 in the base model. The parameter value of an emotion will be given in relation to neutral speech. In Model 1, for instance, the value of the speech rate parameter for joy is 1.1, which means that the rate of happy speech is 10% faster than that of neutral speech; sad speech (0.9) is 10% slower, and angry speech (1.24) is 24% faster than neutral speech. The same logic applies to speech intensity; in all the models, the value of this parameter is lower than 1.0; consequently, emotional speech is 6–45% quieter than neutral speech (see Table 1). The values of the parameters of fundamental frequency (pitch level and pitch range) are represented in semitones in relation to neutral speech. For instance, in Model 2, the mean F0 for joy is 2.5 semitones higher than in neutral speech, and the F0 range is 2.5 semitones wider. The mean F0 of sad speech in Model 2 is 4 semitones lower, and the F0 range is 1.15 semitones narrower.

Table 1. The parametric models of Estonian emotional synthetic speech

Parameter	**Model 1**			
		JOY	SADNESS	ANGER
speech rate		1.1	0.9	1.24
speech intensity		0.9	0.85	0.94
pitch level		1.5	–3.0	–4.0
F0 range		1.5	–0.9	2.1
	Model 2			
	NEUTRAL	JOY	SADNESS	ANGER
speech rate	1.0	1.15	0.8	1.4
speech intensity	1.0	0.85	0.7	0.9
pitch level	0.0	2.5	–4.0	–5.0
F0 range	0.0	2.5	–1.15	2.6
	Model 3			
		JOY	SADNESS	ANGER
speech rate		1.2	0.7	1.56
speech intensity		0.8	0.55	0.86
pitch level		3.5	–5.0	–6.0
F0 range		3.6	–1.4	3.1

The values of the acoustic parameters in Table 1 were used to tune the statistical parametric HTS speech synthesizer for the expression of joy, sadness, and anger. The values of the acoustic parameters, as defined by the acoustic models, were fed into the synthesizer's tuning interface as the variables speech rate, speech intensity, pitch level and F0 range.

5. Method: Evaluation of the test models of emotional synthetic speech using perception tests

5.1. Material

To create the most suitable acoustic model of emotional speech for speech synthesis, we first constructed three test models (see ch. 4), each of which included a set of parameter values characterizing joy, sadness, and anger; the values were either on the optimal level, decreased, or enhanced (see Table 1). We formulated two hypotheses:

- H1: in both the male and the female synthetic voice, the test subjects will best be able to recognize the emotions that were synthesized according to Model 3, and
- H2: the test subjects will recognize the neutral speech synthesized according to the synthesizer's speech model.

In order to test these hypotheses, we designed four perception tests, two for each synthetic voice. The first, Test A, included ten sequences of synthetic speech, each consisting of three sentences,[2],[3] whose acoustic parameters had been modified according to the emotion (either joy, sadness, or anger) and the model (either M1, M2, or M3). One sequence out of the ten remained acoustically neutral (the values of the parameters were set by default by the synthesizer).

The second, Test B, was similar: it included ten sequences of synthetic speech, each consisting of three sentences.[4] In Test B, the sequence began with acoustically neutral speech. But, starting from the second sentence, it either remained neutral or became happy, sad, or angry. The acoustic parameters of each sequence had been modified according to the emotion (either joy, sadness, or anger) and the model (either M1, M2, or M3). One sequence out of the ten remained acoustically neutral (the values of the parameters were set by default by the synthesizer).

2 The text of Test A: "In the evening we went to the restaurant. We ordered the food and waited. When the food was served and we saw it, we all remained speechless."

3 The texts of the synthesized speech in the Tests A and B were constructed using verbal content that would be equally compatible with happy, sad, angry and neutral reading. We wanted the verbal content to have as little influence on the identification of the emotion as possible.

4 The text of Test B: "Wednesday evening I got a call. I was told something I had not expected. Things can take a completely different turn."

The purpose of designing two different tests – Test A and Test B – was to verify whether the change of the emotion, in the course of the speech sequence, contributes to the identification of the emotion.

5.2. Procedure and participants

The perception tests were conducted electronically in the environment of the Estonian Emotional Speech Corpus (Altrov and Pajupuu 2012). The test subjects were asked to listen to the 4 x 10 synthesized speech sequences (2 × male voice (Test A and Test B) and 2 × female voice (Test A and Test B)), and to determine the emotion or the neutrality of each speech sequence. The answer options were joy, sadness, anger, and neutrality. Each test began with an instruction.[5] In Test A, each speech sequence was accompanied by the question "What is the emotion of the sequence?", and in Test B by the question "What is the emotion of the final part of the sequence?" The subjects could listen to each sequence as many times as they wanted, and, if necessary, could modify their earlier answers, or quit the test and continue later.

In order to decide which acoustic features were the most dominant in the emotion models and what their contribution was to the recognition of the emotion, we applied a binary logistic regression to the results of the perception tests.

The test subjects were ten men and ten women between the ages of 30 and 73 (the mean age of the test subjects was 43.3). All were native Estonian speakers. The perception tests were conducted in September 2015.

5 The instruction to test subjects: "*Please listen to ten sequences of synthesized speech and mark for each sequence what emotion do you think it conveys* (version A) */ what emotion do you think its second half conveys* (version B)*. The emotion of the sequence may be either joy, anger, sadness, or neutrality. You can think of it in this way: joy = being pleased; anger = discontent; sadness = melancholy, regret; neutral = no particular emotion. Your opinion will help us know how recognizable the emotion is. You do not have to finish the test now, you can save it and continue later, or modify your earlier answers. Do not forget to save! The "save" button is at the end of the test."*

6. The results of the evaluation of the test models for emotional synthetic speech

6.1. The results of the evaluation of the test models in the case of the male synthetic voice

Table 2 presents the distribution of the test subjects' responses on the recognition of emotions in the male synthetic voice. The target emotion was regarded as recognized correctly only if it had not been confused with any other emotion (no other emotion has a probability exceeding chance) (cf. Altrov and Pajupuu 2015).

Table 2. Confusion matrix (male synthetic voice, Test A/B). Perception of emotions in the male synthetic voice on the basis of the three different test models, responses in percentages. The emotion shaded in grey is the one according to whose parameter values the sound had been modified

Confusion matrix	Response emotions			
Target emotion	Joy	Anger	Sadness	Neutral
Model 1				
Joy	10/30	15/10	45/40	30/20
Anger	15/20	25/10	10/15	50/55
Sadness	10/5	5/0	30/15	55/80
Model 2				
Joy	5/10	35/10	35/70	25/10
Anger	0/10	35/40	10/0	55/50
Sadness	0/5	0/0	30/**50**	70/45
Model 3				
Joy	25/**55**	25/15	35/25	15/5
Anger	5/5	**60**/**50**	10/0	25/45
Sadness	0/0	0/0	35/**80**	65/20
Neutral	10/5	10/5	65/55	15/35

6.1.1. The emotion of joy

Table 2 shows that in Test A, the emotion of joy was not correctly recognized in the male voice. When modelled according to Model 1, joy was taken for sadness[6]; in case of Model 2, it was identified as anger or sadness; and in case of Model 3, it was again taken for sadness.

In Test B, the emotion of joy was correctly recognized in the male voice in Model 3 (55% of respondents); it was not confused with any other emotion or neutral speech. When modelled according to Model 2, the emotion of joy was not recognized and was mistaken for sadness. In Model 1, the result of joy exceeded the level of chance, but even more respondents mistook it for sadness.

When we compare Tests A and B, we see that the emotion of joy was recognized considerably better in Test B, where the switch to the emotion took place in the second part of the speech sequence – i.e. where the listener could compare it to neutral speech. To summarize the results of Tests A and B, the emotion of joy was recognized best in the male voice when modelled according to Model 3, followed by Model 1; joy was least well recognized in male synthetic voice according to Model 2. The emotion of joy was most often confused with sadness.

6.1.2. The emotion of anger

Table 2 shows that in Test A, the emotion of anger was correctly recognized in the male voice according to Model 3 (by 60% of the respondents), without it being confused with other emotions or neutral speech. When modelled according to Model 2, the result for the emotion of anger did exceed the probability of chance, but there were more respondents who mistook it for neutral speech. Anger was also not recognized when modelled according to Model 1, at which point it was also mistaken for neutral speech.

In Test B, like in Test A, anger was recognized in the male voice according to Model 3 (by 50% of the respondents). Here, the option "neutral" also passed the level of chance. When modelled according to Model 2, the result for anger did exceed the probability of chance, but more respondents mistook it for neutral speech. When modelled according to Model 1, anger was not recognized and was mistaken for neutral speech.

6 The result exceeds the probability of chance, which is 25%.

When we compare Tests A and B, we see that the emotion of anger was better recognized in Test A, where the whole sequence conveyed the same emotion. Summarizing the results of Tests A and B, we can say that the emotion of anger was recognized best in the male voice according to Model 3, followed by Model 2; anger was least well recognized in the male synthetic voice according to Model 2. The emotion of anger was most often confused with neutral speech.

6.1.3. The emotion of sadness

Table 2 shows that in Test A, the emotion of sadness was not correctly recognized in the male voice. The result for the emotion of sadness when modelled according to Model 3 did exceed the level of chance, but more respondents mistook it for neutral speech. In case of Models 1 and 2, the result of sadness also passed the level of chance, but more respondents mistook it for neutral speech.

In Test B, the emotion of sadness was recognized in the male voice according to Model 3 (by 80%) and Model 2 (by 50%). In the case of Model 3, it was not confused with other emotions or neutral speech, whereas in the case of Model 2, the option "neutral" also passed the level of chance. When modelled according to Model 1, sadness was not recognized and was mistaken for neutral speech.

Comparing the results of Tests A and B, we see that the emotion of sadness was recognized better in test B, where the emotion appeared in the second part of the speech sequence – i.e. where the listeners had the possibility of comparing it with neutral speech. To sum up the results of Tests A and B, we can claim that sadness was recognized best in the male voice according to Model 3, followed by Model 2; sadness was least well recognized in the male voice according to Model 1. Sadness was most often confused with neutral speech.

6.1.4. Neutral speech

Neutral speech was synthesized according to the statistical parametric speech model of the speech synthesizer, the values of the acoustic parameters of neutral speech being independent of the results of the study of human speech. The results of the perception test showed that neutral speech was not recognized in the male voice in either Test A or in Test B (see Table 4). In Test A, it was mistaken for sadness, and although in Test B the recognition percentage of neutral speech

exceeded the level of chance, there were more subjects who mistook it for sadness.

When we compare the results of Tests A and B, we see that the result was 50% better in Test B. Neutral speech was most often confused with sadness.

6.2. The results of the evaluation of the test models in case of the female synthetic voice

Table 3 presents the distribution of the test subjects' responses on the recognition of emotions in the female synthetic voice. The target emotion was regarded as recognized correctly only if it had not been confused with any other emotion (no other emotion has a probability exceeding chance) (cf. Altrov and Pajupuu 2015).

Table 3. Confusion matrix (female voice, Test A/B). Perception of emotions in the female synthetic voice on the basis of the three different test models, responses in percentages. The emotion shaded in grey is the one according to whose parameter values the sound had been modified

Confusion matrix	Response emotions			
Target emotion	Joy	Anger	Sadness	Neutral
Model 1				
Joy	**40**/5	25/20	20/55	15/20
Anger	45/25	25/25	10/5	20/45
Sadness	10/10	10/0	30/30	50/60
Model 2				
Joy	25/10	5/35	40/50	30/5
Anger	30/5	**45**/**50**	0/30	25/15
Sadness	0/10	0/0	**60**/35	40/55
Model 3				
Joy	**30**/10	20/40	25/50	25/0
Anger	15/20	**65**/**60**	5/10	15/10
Sadness	5/10	5/0	**50**/**75**	40/15
Neutral	15/25	10/5	25/30	**50**/**40**

6.2.1. The emotion of joy

Table 3 shows that the emotion of joy was correctly recognized in the female voice according to Models 1 (by 40%) and 3 (by 30%). In neither was it confused with other emotions or with neutral speech. In the case of Model 2, joy was not recognized, being mistaken either for sadness or for neutral speech.

In Test B, joy was not correctly recognized in the female voice. When modelled according to Model 3, joy was mistaken either for sadness or for anger. The same goes for Model 2. When modelled according to Model 1, joy was mistaken for sadness.

A comparison of the results of Tests A and B shows that joy was considerably better able to be recognized in Test A, where the whole sequence conveyed the same emotion. Summarizing the results of Tests A and B, we can say that joy was recognized best in the female voice according to Model 1, followed by Model 3; joy was least well recognized in the female synthetic voice according to Model 2. Joy was most often confused with sadness.

6.2.2. The emotion of anger

Table 3 shows that in Test A, the emotion of anger was correctly recognized in the female voice according to Models 3 (by 65% of respondents) and 2 (by 45% of respondents). In the case of Model 3, anger was not confused with other emotions or with neutral speech; in the case of Model 2, the option "joy" also exceeded the probability of chance. When modelled according to Model 1, anger was not recognized and was instead mistaken for joy.

In Test B, anger was also recognized according to Models 3 (by 60% of respondents) and 2 (by 50% of respondents). In the case of Model 3, anger was not confused with other emotions or with neutral speech; in the case of Model 2, the option "sadness" also exceeded the level of chance. When modelled according to Model 1, anger was not recognized and was mistaken for neutral speech.

Comparing the results of Tests A and B, we see that anger was equally well recognized in both. To sum up the results of Tests A and B, we can say that the emotion of anger was recognized best in the female voice according to Model 3, followed by Model 2; anger was least well recognized according to Model 1. Anger was most often confused with joy.

6.2.3. The emotion of sadness

As can be seen from Table 3, in Test A, the emotion of sadness was correctly recognized both according to Model 2 (by 60% of respondents) and Model 3 (by 50% of respondents). For both models, the option "neutral" also exceeded the probability of chance. In the case of Model 1, sadness passed the level of chance, but more respondents took it for neutral speech.

In Test B, sadness was correctly recognized in the female voice according to Model 3 (by 75% of respondents). It was neither confused with the other emotions nor with neutral speech. In the case of Model 2, sadness did exceed the probability of chance, but there were more subjects who mistook it for neutral speech. The same was true for Model 1.

A comparison of the results of Tests A and B shows that sadness was equally well recognized in both. Summarizing the results of Tests A and B, we can claim that sadness was recognized best in the female voice according to Model 3, followed by Model 2; sadness was least well recognized from Model 1. Sadness was most often confused with neutral speech.

6.2.4. Neutral speech

Neutral speech was synthesized according to the synthesizer's existing model. Table 3 shows that in both tests, neutral speech was correctly recognized in the female voice. In Test A, it was not confused with the emotions; in Test B, the option "sadness" also exceeded the level of chance.

A comparison of the results of Tests A and B shows that neutrality was better recognized in Test A. In the case of the female voice, neutral speech was not confused with the emotions.

6.3. The overall analysis of the results of the perception tests and the evaluation of the test models

The respondents' average recognition rate of the emotions across all the models was 35% (i.e. 14 correct answers out of 40). The overall recognition level was 32% in the case of the male synthetic voice, and 38% in the case of the female voice (both exceeded the level of chance). We hypothesized that listeners would be able to recognize emotions best in the speech synthesized according to Model 3. This hypothesis was

borne out; the average recognition rates were highest for M3: sadness was recognized in 60% of the cases, anger in 59%, and joy only in 30%. The recognition level of neutral speech was 35%. The average recognition percentages of the emotions were relatively similar to the results that were obtained in the synthesis experiment of Catalan emotional speech (Iriondo et al. 2004). In their perception tests, too, joy was the least well recognized emotion (being recognized in only 30% of the cases), sadness was almost always recognized (in 90% of the cases), and anger was recognized, on average, in 48% of the cases.

In order to decide which acoustic features were the most dominant in the emotion models, and what their contribution was to the recognition of the right emotion, we applied a binary logistic regression to the results of the perception test. The analysis also allowed us to indirectly evaluate whether the choice of the value of each parameter in all the emotion models was correct or not. A further goal was to determine whether the acoustic features had the same effect in the male and female voice regarding the expression of emotions in synthetic speech, or whether there were significant differences.

The dependent variable of the binary logistic regression was the (non-)identification of the emotion or of the neutral speech (TRUE *vs.* FALSE), and the argument features were the values of the speech rate, voice intensity, pitch and pitch range parameters. Table 4 presents the results of the logistic regression analysis performed on the data of the perception test.

Table 4. Logistic regression analysis of emotion recognition in the perception tests (significant values in bold)

Parameter	Est.	S.E.	Z	p-value
Constant	–2.331	0.801	–2.909	**0.004**
Speech_rate	–5.629	1.927	–2.920	**0.003**
F0_level	–0.047	0.067	–0.698	0.485
Speech_intensity	5.443	1.289	4.222	**<0.001**
F0_range	3.782	1.444	2.620	**0.009**
Overall model fit				
Chi-Square			77.758	
p-value			**<0.001**	
Total correct classifications			0.688	

The speech rate, voice intensity and pitch range all played a significant role in the recognition of the correct emotion in the perception test. Only the pitch register, or F0 level, did not prove to be a significant parameter in the perception of emotions. The logistic regression model was statistically significant, although its classification power was not very high (69%).

Both in case of the male and the female synthetic voice, the probability of the right emotion being recognized was most strongly influenced by the appropriate intensity and F0 range. Table 5 shows the odds ratio estimates as to by how many times the appropriate value of a parameter affects the perception of the right emotion. For instance, in the case of the male synthetic voice, the appropriate intensity or loudness of the speech increases the perception of the right emotion by as much as 257 times, and the appropriate F0 range by 87 times. In the case of the female synthetic voice, the appropriate F0 range increases the probability of the right judgement being made by 134 times, and the appropriate loudness by 47 times. However, these estimates must be treated with caution: first, because they are based on a single series of perception tests, and second, because they are more accurately interpreted as characterizing the correctness of the choice of the parameter values in the emotion models. Most likely the values of the speech intensity and the F0 range parameters were chosen relatively correctly in all the models, given that they were significant in the overall model and contributed significantly to the perception of the right emotion, and consequently also to the generation of the right emotion both in the male and female synthetic voice. The speech rate was a significant parameter in the logistic regression model, but its contribution to the recognition of the emotions was marginal, being limited to a few per cent; only in the recognition of sadness did it play a more significant role. The speech rate values of anger and joy still need to be specified. The pitch register, or F0 level, however, needs to be assigned appropriate values in all the models, since it turned out to be an insignificant parameter in the overall regression model.

Table 5. Odds ratio estimates of speech intensity and F0 range for male and female synthetic voices in the recognition of emotions

Synthetic voices	Odds ratio estimates	
	Speech intensity	F0 range
Male voice	256.7	87.0
Female voice	47.1	134.4

7. Discussion

We hypothesized concerning the test models of three basic emotions that, both in the male and the female synthetic voice, the test subjects would recognize best the emotions that have been synthesized according to the model with enhanced parameter values, M3, as well as the neutral speech synthesized according to the synthesizer's own speech model. The test results of the models showed that the hypothesis was partly right: both in the male and the female synthetic voice, listeners recognized best the emotions that had been synthesized according to M3 (an exception is the emotion of joy synthesized on the basis of the female voice, which was best recognized according to Model 1). In case of neutral speech, the results diverged: neutrality was not correctly recognized in the male voice, being instead mistaken for sadness. In the female synthetic voice, neutrality was correctly recognized.

The comparison of Tests A and B shows that in the case of the male synthetic voice, the synthesized emotions were better recognized in Test B, where the speech sequence began with acoustically neutral speech and subsequently, starting from the second sentence, either became happy, sad, or angry, or remained neutral. In the case of the female synthetic voice, anger and sadness were equally well recognized in both tests, whereas joy and neutrality were better recognized in Test A, where the whole sequence conveyed the same emotion or was neutral. Consequently, in the case of the male synthetic voice, the recognition of emotions was facilitated by contrast, a switch from neutral to emotional speech.

On the basis of the test results of the models, we can choose the preferred test models of basic emotions for parametric synthesis (see table 6).

Table 6. The preferred test models of three basic emotions for parametric synthesis

Male synthetic voice		Female synthetic voice	
joy	M3	*joy*	M1
anger	M3	*anger*	M3
sadness	M3	*sadness*	M3

For neutral speech, the Estonian speech synthesizer has a speech model for each synthetic voice, which has been trained on a corpus of neutral speech. However, since the evaluation results of the test models showed that neutrality was not correctly recognized in the male synthetic voice, the neutral speech model for this voice will need to be further fine-tuned.

The synthesizer was least successful at conveying joy in synthesized speech; in the case of M1 and M2, the average recognition percentage of joy was below the level of chance, and in the case of M3, exceeded it only marginally (30%). Figure 1 shows that the test models of joy and anger were located relatively close to each other in the acoustic space of the speech rate and speech intensity. However, the confusion matrices in Tables 4 and 5 show that joy was taken for anger only in 21% of the cases. It was much more frequently (in 41% of the cases) perceived as sadness. Sadness was the most recognizable emotion: in case of M3, it was recognized in 60% of the cases and in some test series in up to 80%. Consequently, the parameter values of sad speech in the four-dimensional acoustic parameter space are relatively well established. Hence the parameter values of happy speech must be shifted as far as possible from the region of sad speech; for instance by increasing the intensity of the speech and raising the pitch level. The logistic regression analysis of the perception tests (see 6.3) also showed that in the current models, the pitch level does not play a significant role. Consequently, in order to adjust the parameter values of the models of anger and, in particular, joy, we should first specify the characteristic registers of these emotions.

8. Conclusion

The goal of this study was to conduct modelling experiments in order to ascertain whether, and to what extent, emotions' characteristic acoustic parameter values that have been identified on the basis

of human speech are able to generate emotions in synthetic speech. For three emotions (joy, sadness and anger), three different test models were constructed and then evaluated by test subjects in perception tests. Among the emotions, sadness was the one that was recognized most frequently in synthetic speech: in 60% of the cases, on average. Anger was identified, on average, in 59% of the cases. The emotion that was least well conveyed by the synthesizer was joy: its mean recognition percentage was below the level of chance in the case of M1 and M2, and only marginally above it in the case of M3 (30%).

The test results showed that the emotion model that performs best in the Estonian speech synthesizer is M3, in which the values of the acoustic parameters were enhanced by 15% in comparison with the results obtained in the study of human speech. The only exception is the emotion of joy in the female synthetic voice, which is best synthesized according to Model 1 (the model with decreased values).

The overall logistic regression analysis of the results of the perception test showed that the values of the speech intensity and F0 range parameters were the ones that played the most significant role in the models that had been constructed for the basic emotions. Speech rate was a significant parameter in the logistic regression model, but its contribution to the recognition of the emotions was marginal. The values of the F0 level, or the pitch register, parameter, however, need to be adjusted in all the models, because this parameter was not significant in the overall regression model. The parameters of the model of happy speech need to be adjusted most, but the mean F0 of angry and sad speech also requires some revision. These first attempts to generate basic emotions in the Estonian parametric speech synthesis will serve as a starting point for a further development of the emotion models.

Acknowledgements

This work was supported by the institutional research funding IUT35-1 of the Estonian Ministry of Education and Research.

Addresses:

Kairi Tamuri
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
E-mail: kairi.tamuri@eki.ee

Meelis Mihkla
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
E-mail: meelis.mihkla@eki.ee

References

Altrov, Rene (2013) "Aspects of cultural communication in recognizing emotions". *Trames* 17, 159–174.

Altrov, Rene and Hille Pajupuu (2012) "Estonian Emotional Speech Corpus: theoretical base and implementation". In Laurence Devillers, Björn Schuller, Anton Batliner, Paolo Rosso, Ellen Douglas-Cowie, Roddy Cowie, and Catherine Pelachaud, eds. *4th international workshop on corpora for research on emotion sentiment & social signals (ES3)*, 50–53. Istanbul.

Altrov, Rene and Hille Pajupuu (2015) "The influence of the language and culture on the understanding of vocal emotions". *Journal of Estonian and Finno-Ugric Linguistics. Special issue "Aspects of Speech Studies"*, xx–xx.

Audibert, Nicolas, Véronique Aubergé, and Albert Rilliard (2005) "The prosodic dimensions of emotion in speech: the relative weights of parameters". *Proceedings of the 9th international conference on speech communication and technology* (INTERSPEECH 2005), 525–528. Lisbon, Portugal.

Ekman, Paul (1992) "Are there basic emotions?" *Psychological Review* 99, 3, 550–553.

Iriondo, Ignasi, Francesc Alías, Javier Melenchón, and M. Angeles Llorca (2004) "Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis". In Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp, eds. *Affective dialogue systems: tutorial and research workshop; ADS 2004*, 197–208. Berlin et al.: Springer.

Lorenzo-Trueba, Jaime, Roberto Barra-Chicote, Ruben San-Segundo, and Javier Ferreiros (2015) "Emotion transplantation through adaptation in HMM-based speech synthesis". *Computer Speech & Language* 34, 1, 292–307.

Mihkla, Meelis, Indrek Hein, Indrek Kiissel, Artur Räpp, Risto Sirts, and Tanel Valdna (2013) "Subtiitrite helindamine – kas, kuidas, kellele ja milleks?". [Spoken subtitles – if, how, for whom and why?] *Keel ja Kirjandus* 11, 819–828.

Mihkla, Meelis, Indrek Hein, Mari-Liis Kalvik, Indrek Kiissel, Risto Sirts, and Kairi Tamuri (2012) "Estonian speech synthesis: applications and challenges / Sintez reči èstonskogo jazyka: primenenie i vyzovy". In Alexander E. Kibrik, ed. *Computational linguistics and intellectual technologies,* papers from the annual international conference "Dialogue", 443–453. Moskva: RGGU.

Paulmann, Silke and Ayse K. Uskul (2014) "Cross-cultural emotional prosody recognition: evidence from Chinese and British listeners". *Cognition and Emotion* 28, 2, 230–244.

Scherer, Klaus (2013) "Vocal markers of emotion: comparing induction and acting elicitation". *Computer Speech & Language* 27, 1, 40–58.

Tamuri, Kairi (2010) "Kas pausid kannavad emotsiooni?". [Do the pauses in read text carry emotion?] *Eesti Rakenduslingvistika Ühingu aastaraamat* 6, 297–306.

Tamuri, Kairi (2012a) "Kas formandid peegeldavad emotsioone?". [Do formants speak of emotions?] *Eesti Rakenduslingvistika Ühingu aastaraamat* 8, 231–243.

Tamuri, Kairi (2012b) "Intensity of Estonian emotional speech". In *Human language technologies – the Baltic perspective – proceedings of the fifth international conference Baltic HLT 2012*, 238–246. Amsterdam: IOS Press.

Tamuri, Kairi (2015) "Fundamental frequency in Estonian emotional read-out speech". *Journal of Estonian and Finno-Ugric Linguistics* 6, 1, 9–21.

Tamuri, Kairi and Meelis Mihkla (2012) "Emotions and speech temporal structure". *Linguistica Uralica* 3, 209–217.

Yamagishi, Junichi, Koji Onishi, Takashi Masuko, and Takao Kobayashi (2005) "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis". *IEICE Transactions on Information and Systems* E88-D, 3, 503–509.

Zen, Heiga, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black, and Keiichi Tokuda (2007) "The HMM-based speech synthesis system (HTS) version 2.0". 6th ISCA workshop on speech synthesis, 294–299. Bonn, Germany.

Kokkuvõte. Kairi Tamuri ja Meelis Mihkla: Põhiemotsioonide väljendusvõimalused eestikeelsel parameetrilisel kõnesünteesil. Uurimistöö eesmärk oli läbi viia modelleerimiseksperimente kolme põhiemotsiooni (rõõmu, kurbuse ja viha) väljendamiseks eestikeelsel parameetrilisel kõnesünteesil nii mees- kui ka naissünteeshääle baasil. Selleks koostati iga emotsiooni kohta kolm erinevat katsemudelit, mida lasti katseisikutel tajutestidel hinnata. Katsemudelite aluseks oli inimkõne põhjal määratud põhiemotsioonidele omased parameetrite väärtused. Emotsioonidest tunti sünteeskõnes kõige paremini ära kurbuse-emotsioon ning kõige halvemini rõõmu-emotsioon. Testitulemused näitasid, et kui meessünteeshääle puhul töötas kõigi kolme emotsiooni puhul kõige paremini võimendatud väärtuste mudel, siis naissünteeshääle puhul vajasid erinevad emotsioonid erinevaid mudeleid: rõõmu väljendamiseks sobis kõige paremini vähendatud väärtuste mudel, kurbuse ja viha väljendamiseks võimendatud väärtuste mudel. Tajutestide tulemusi analüüsiti logistilisel regressioonil, et teha kindlaks üksikute akustiliste parameetrite olulisus ja osakaal emotsioonimudelites ning parameetrite väärtuste korrigeerimisvajadused.

Märksõnad: eesti keel, emotsioonid, kõnesüntees, akustiline mudel, kõnetempo, intensiivsus, põhitoon