Altrov, Rene and Pajupuu, Hille. 2008. “The Estonian emotional speech corpus: Release 1.” *Proceedings of the Third Baltic Conference on Human Language Technologies: The Third Baltic Conference on Human Language Technologies*, František Čermak, Rūta Marcinkevičienė, Erika Rimkutė, Jolanta Zabarskaitė (eds.), 9-15. Vytauto Didžiojo Universitetas, Lietuviu Kalbos Institutas, Vilnius.

HLT' 2007: THE ESTONIAN EMOTIONAL SPEECH CORPUS: RELEASE 1

Rene Altrov, Hille Pajupuu
Institute of the Estonian Language (Estonia)

Abstract

The Estonian emotional speech corpus (EEKK) is being created in the framework of the National Programme for Estonian Language Technology in the Institute of the Estonian Language. The corpus contains recordings of read speech sentences expressing anger, joy and sadness, and neutral sentences. The corpus serves two objectives: 1) to form an acoustic basis of corpus-based emotional text-to-speech synthesis; 2) to provide a reliable database for studying emotions rendered by speech. The underlying principle of the corpus is that emotions can be relatively well recognised in natural, non-acted speech which is a precondition for synthesising natural speech (see Iida et al. 2003). The reliability of the corpus is ensured by perception tests: each corpus sentence is provided with perception test results on the recognisability of an emotion. The corpus is now in the testing stage. The present article gives an overview of the stages of corpus creation and results of perception tests.[1]

Keywords: speech corpus, emotions, natural speech, perception tests, Estonian.

1. Introduction

We started working on the Estonian emotional speech corpus in 2006 in the framework of the National Programme for Estonian Language Technology (see Altov in print; 2007). The corpus serves two objectives: 1) to form an acoustic basis of corpus-based emotional text-to-speech synthesis; 2) to provide a reliable database for studying emotions rendered by speech. At first, we decided to include in the corpus sentences expressing the three basic emotions – anger, joy and sadness – and neutral sentences, reserving an opportunity to expand the corpus with sentences rendering other emotions or emotion-like conditions (see Scherer 2000). Because the speech synthesisers presently available in Estonia have been designed to read written texts and not to develop spontaneous conversations with humans, the corpus contains read-speech

[1] This work was completed within the framework of the National Programme for Estonian Language Technology.

sentences. An essential aspect in creating the corpus was not to dictate to the readers of texts which emotions sentences should express but to allow them to decide on how they read each sentence. Listeners are actually the ones who decide which emotion each sentence renders: in perception tests they were asked to listen to the recordings and decide whether the sentences they heard expressed joy, sadness, anger, or were neutral in nature. In the corpus, all sentences are presented with the results of the perception test. In a query, a corpus user can determine the recognition percentage of emotions contained in sentences. The higher the recognition percentage, the more certain it is that the sentence carries the desired emotion. As the corpus is in every way open for expansion, the recognition percentages of sentences may change when more listeners and repetitive perception tests are added to the sample. By default, the recognition percentage of an emotion is currently 51%. This means that, with more listeners, borderline cases may end up in a different emotion category.

Relying on earlier emotional speech research and the experience of the creators of the existing emotional speech corpora (see e.g. Douglas-Cowie et al. 2003), we considered the following aspects important in creating our corpus. First, we decided not to use acting as acted emotions tend to be stereotypic and exaggerated in nature and different from real-life communication, which is why readers are unable to decode the real meaning of emotions (Douglas-Cowie et al. 2003; Scherer 2003). Research has shown that readers are relatively good at recognising emotions also in natural speech (Iida et al. 2003). Therefore, as our aim is as natural speech synthesis as possible, we decided to use non-acted, natural emotions.

Second, the reader's voice needs to be pleasant because it will constitute the voice material of the synthesiser. The pleasantness of voice was again determined by listeners. The reader must have good articulation and empathic abilities. Empathic readers are better at rendering the emotions contained in a text (see, Baron-Cohen, Wheelwright 2004). Third, recognising an emotion by listening to a recording and not seeing the speaker is strongly culture-dependent: the longer one has lived in a culture, the better one is at recognising emotions by listening to an audio only (Toivanen et al. 2004). Therefore our listeners are over 30-year old Estonians whose main language of education has been Estonian. Fourth, empathic people are better at recognising recorded emotions than unempathic ones. That is why our listeners were subjected to the Estonian translation version of the empathy test designed by Baron-Cohen and Wheelwright (2004). Relying on test results, we excluded unempathic listeners.

Next, we will describe corpus creation and verifying emotions by perception tests.

2. Material and Method

Our starting point was that each text renders an emotion or a neutral message. We took texts from the media and adapted them for reading out loud. We chose 26 text passages (see example 1):

(1) *Ma ei räägi enam olukorrast, kus liinilt kaob paar bussi, vaid et päevas jääb ära 20-30 väljumist! Mida peavad tegema need inimesed, kes autot pole? Need ei pääse ju üldse liikuma! Kord on Harjumaa bussiliikluses täiesti käest ära! Viivitamatult tuleks leida uus vedaja ja lõpetada praeguse bussifirma monopoolsus. Sellele on ju pidevalt probleeme – kord on neil bussid katki ja puuduvad juhid, siis kurdavad vähese dotatsiooni üle.*

[I'm not talking about a situation where a line loses a couple of busses; instead 20-30 buses are cancelled every day! What should people without cars do? They won't be able to move at all! There is a chaos in bus traffic in Harjumaa! We need to immediately find a new carrier and put an end to the monopoly of the current one. It is always having problems – either buses are broken and drivers are missing or it is complaining about small subsidies.]

At first, 10 people were asked to read the text passages silently and decide which emotions the passages contained. Our treatment of emotions was wider than the stereotypic understanding. For example, in our culture *anger* stereotypically means rage, whereas in our treatment anger also included resentment, irony, repulsion, contempt and malice. Joy included gratitude, happiness, pleasantness and admiration. Sadness included loneliness, disconsolation, concern, despair; neutrality meant that the text rendered no particular emotions. Test subjects were also given such an extended description of emotions.

We excluded one of the 26 text passages as readers were unable to recognise any emotions in it.

We then selected the readers. We asked three women to read one and the same text passage, and then organised a voice test. We asked listeners to assess the pleasantness of the readers' voices. The instruction was: *Listen to three recorded passages and give your personal opinion about each voice.* The results are given in Table 1.

Table 1. Listeners' assessment of readers' voices. Values are given in percentages

Statement and assessment / **Voice**	**I like this voice.**			
	Not true	**Not quite true**	**Almost true**	**True**
Voice 1	**50.0**	**45.2**	4.8	0.0
Voice 2	17.1	**41.5**	**36.6**	4.9
Voice 3	4.9	12.2	**48.8**	**34.1**

We then recorded texts passages read by the two readers with the most pleasant voices (voice 2 and voice 3). We instructed the readers: *Read the text so that you render its mood.* Readers could decide how to interpret each text passage.

Recording was done in a sound studio of the Institute of the Estonian Language, using Edirol-R09 (sampling frequency 44.1 kHz, 16 bits, stereo). The recorded speech waves were segmented into sentences with Sony SoundForge 8.0. Uneven breaks between passages were replaced with 0.2 second silence breaks. Such labelled speech waves are kept in the database together with texts and reader data.

We compiled two perception tests (separately for each reader). The underlying principle was that the content of two successive sentences could not form a logical sequence. Perception test subjects had to listen to isolated sentences without seeing the text and decide which emotion the sentences contained.[2] The options were three

[2] Perception test instructions: Although people recognise emotions mostly by relying on both voice and facial expression, they can also be verified by simply listening to the speaker's voice.

Although in normal speech speakers do not exaggerate emotions the way actors do, it is quite easy to recognise them. However, normal speech rarely contains clear emotions. When you are listening to the following sentences which are cut-outs from recorded texts, we are asking you to mark the emotion that characterises each sentence most.

You can choose between the three main emotions – joy, sadness, anger – and neutral speech.

Footnote continued on the next page

emotions – sadness, anger, joy – or neutral speech. In total, subjects were asked to listen to 173 sentences (1473 tokens). Perception tests were carried out using a web-based test environment. Each subject was given a user name and password to access the system. We asked the subjects' gender, education, nationality, native language, language of education and age.

To do the perception test, subjects had to have internet access, a computer with an audio card and loudspeakers or headphones. It was also important that they had enough time to do the test. For that, we contacted the potential candidates earlier to explain the content and structure of the test and the time needed to do it. Only those candidates who agreed to do the test were given access to the system.

Each sentence was listened to by at least 34 Estonians who were over 30 and had good empathic abilities. As our aim was to screen out sentences in which emotions could be recognised by listening an audio recording only, we had to eliminate any side-effects of the text to recognition of emotions. To do that, we organised another test: we asked 14 subjects who had not participated in the perception test to decide on the emotion or neutrality of each sentence by reading the text (and not listening to the audio).

3. Results of perception tests

Emotions determined by listening or reading only did not always coincide. This led to the establishment of two categories:

1. Emotion is rendered by audio text only.
   - Listeners cannot determine an emotion in written text but can do it when listening to a text. See Table 2, 1st sentence.
   - One emotion (e.g. *anger*) is recognised in written text, a different one (e.g. *sadness*) when listening to the text. See Table 2, 2nd sentence.
2. Recognition of an emotion in the perception test may have been influenced by written text.
   - One and the same emotion (e.g. *anger*) is recognised in written text and when listening to the text. See Table 2, 3rd sentence.

The audio texts falling to the first category constitute the acoustic basis of synthesised speech. Recognition of emotions in sentences falling to the second category is somewhat influenced by text, which is why those sentences should be viewed separately to find out what (lexis, syntax, etc) enables to recognise the emotion contained in the text.

As a result of a perception test, of all sentences, 79 were perceived as expressing anger, 20 expressing sadness, 3 expressing joy and 25 as neutral ones. For more details, see Table 3.

To make the decision process easier, take a look at the following list of emotions contained in the main emotions: Joy = gratitude, happiness, pleasantness, admiration; Anger = resentment, irony, repulsion, contempt, malice, rage; Sadness = loneliness, disconsolation, concern, despair; Neutral = ordinary speech without any particular emotions

Listen to each sentence and, relying on the voice, try to decide which emotions they contain.

There are no right and wrong answers. On the contrary, your opinion helps us to find out how recognisable different emotions are. You do not have to complete the test at once. Feel free to save the sentences, take breaks or change your earlier opinions. Do not forget to save!

There were 46 sentences in which listeners were not able to recognise any emotion or neutrality. Such sentences will be subjected to further testing. In upcoming tests, readers will no longer be given a list of emotions to choose from. Instead they can decide on their own which emotion each sentences expresses. This way we can establish new emotions which have not been included in the corpus yet.

Table 2. Classification principles of emotion sentences. Values are given in percentages

1st sentence: *Ootaks valitsuse vastust, aga valitsust ju ei ole. [I'd expect the government's response, but there is no government.]*	**neutral**	**joy**	**not sure**[3]	**sadness**	**anger**	**comment**	**sentence type in corpus**
by reading	0.0	9.1	0.0	45.5	45.5	unable to determine emotion	*anger*, no content influence, rendered by audio only
by listening	0.0	0.0		25.0	**75.0**	recog-nised as anger	
2nd sentence: *Katkev internet, tarduv ja ruuduline telepilt, kättesaamatu tehnilise toe telefon, mitte töötav koduleht…! [Abrupt internet connection, freezing and checked TV-screen, busy technical support line and non-working homepage...!]*	**neutral**	**joy**	**not sure**	**sadness**	**anger**	**comment**	**sentence type in corpus**
by reading	0.0	0.0	0.0	18.2	**81.8**	recog-nised as anger	*sadness*, no content influence, rendered by audio only
by listening	10.0	6.7		**63.3**	20.0	recog-nised as sadness	
3rd sentence: *Igasugustel peksta ma ennast ka ei lase! [I'm not going to have just anyone beat me!]*	**neutral**	**joy**	**not sure**	**sadness**	**anger**	**comment**	**sentence type in corpus**
by reading	0.0	0.0	7.1	7.1	**85.7**	recog-nised as anger	*anger*, content influence
by listening	3.1	0.0		6.2	**90.6**	recog-nised as anger	

Table 3. Classification of emotion sentences based on reception and reading tests

Perception tested sentences in the corpus	**Emotion rendered by audio only**	**Possible text influence on recognition of emotion**	**Total**	**% of all test sentences**
sentences expressing anger	28 (192 tokens)	51 (434 tokens)	79	45.7
sentences expressing sadness	17 (173 tokens)	3 (12 tokens)	21	11.6
sentences expressing joy	3 (11 tokens)	-	3	1.7
neutral sentences	16 (147 tokens)	9 (81 tokens)	25	14.4
unable to determine	46 (427 tokens)		46	26.6

[3] The category "not sure" was added in case subjects find it hard to decide one particular emotion in a sentence and feel that the emotion rather depends on how the sentence is read.

4. Queries

In the Estonian emotional speech corpus sentences are stored as segmented and labelled speech waves, text and related results of perception tests and reading tests.

The corpus can currently be searched for sentences expressing anger, joy, sadness and neutral sentences. Each sentence is provided with the emotion (or neutrality) recognition percentage. Sentences can also be searched by the recognition percentage of emotions (51% by default). Queries can be restricted to: 1) only those sentences where emotion is rendered by audio; 2) only those sentences where recognition of emotions may have been influenced by text. Sentences are displayed as text and can be listened to by clicking on them (see Figure 1).

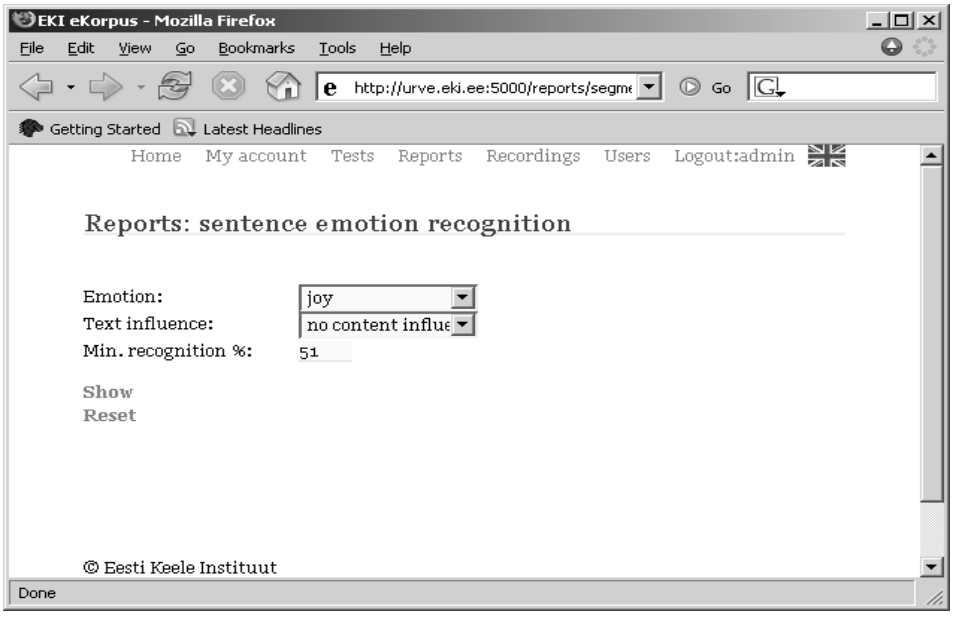

Figure 1. The Estonian Emotional Speech Corpus: Queries

5. Conclusion

Our emotional speech corpus is a web-based application that uses freeware: Linux, PostgreSQL, Python, Praat, and is intended for public domain use. The corpus is not ready yet; we are trying to build an open platform which will satisfy all our research needs. For now our system supports preparing, executing and analysing automatic perception tests. The next step is segmentation and labelling of the speech wave and making the queries of information possible by integrating the Praat programme with our corpus (Boersma, Weenink 2007). We will continue expanding our corpus with sentences containing the main emotions. We are also working on sub-classification of emotions, e.g. differentiating between resentment, malice, irony, etc in anger emotion. We are planning to enlarge the corpus also with neutral sentences to ensure a sufficient acoustic basis for speech synthesis (see Nurk et al. in this Vol.)

6. References

Altrov, Rene. 2007. Emotsionaalse kõne korpuse loomine eesti keele tekst−kõne sünteesi jaoks. Tekstimaterjali evalvatsioon *viha* näitel. Magistritöö. Tartu Ülikooli filoloogiateaduskond. Eesti ja soome-ugri keeleteaduse osakond. http://dspace.utlib.ee/dspace/bitstream/10062/2739/1/altrov_rene.pdf

Altrov, Rene. In print. Eesti emotsionaalse kõne korpus: teoreetiline tagapõhi. [The Estonian emotional speech corpus: Theoretical background.] Keel ja Kirjandus.

Baron-Cohen, Simon, Wheelwright, Sally. 2004. The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism and normal sex differences. Journal of Autism and Developmental Disorders, vol. 34 (2), 163–175.

Boersma, Paul; Weenink, David. 2007. Praat: doing phonetics by computer (Version 4.6.31) [Computer program]. Retrieved October 8, 2007, from http://www.praat.org/

Douglas-Cowie, Ellen; Campbell, Nick; Cowie, Roddy; Roach, Peter. 2003. Emotional speech: towards a new generation of databases. *Speech Communication*, 40 (1-2), 33–60.

Iida, Akemi, Campbell, Nick, Higuchi, Fumito, Yasumura, Michiaki. 2003. A corpus-based speech synthesis system with emotion. *Speech Communication*, vol. 40, 161–187.

Nurk, Tõnis; Mihkla, Meelis, Kiissel, Indrek, Piits, Liisi. In this Vol. Development of unit selection TTS system for Estonian. The Third Baltic Conference on Human Language Technologies. Kaunas: Vytautas Magnus University and Institute of Lithuanian Language.

Scherer, Klaus R. 2000. Emotion effects on voice and speech: paradigms and approaches to evaluation. Paper presented at the *ISCA w/s on Speech and Emotion*. Newcastle, Northern Ireland. 5.–7.9.2000, 39-44.

Scherer, Klaus R. 2003. Vocal communication of emotion: A review of research paradigms, *Speech Communication* vol. 40, 227–256.

Toivanen, Juhani, Väyrynen, Eero, Seppänen, Tapio. 2004. Automatic discrimination of emotion from spoken Finnish. Language and Speech 47 (4), 383–412. http://emosamples.syntheticspeech.de (16.10.2007).

RENE ALTROV, MA, Researcher Extraordinary, Institute of the Estonian Language, Tallinn, Estonia. Doctoral studies of general linguistics and Finno-ugric languages at the University of Tartu. Fields of research: developing linguistic corpora (corpus of emotional speech); intercultural communication (proxemics, contextuality of language). E-mail: rene.altrov[at]eki.ee.

HILLE PAJUPUU, PhD in philology, Scientific Secretary, Institute of the Estonian Language; docent, The Estonian Information Technology College, Tallinn, Estonia. Fields of research: speech acoustics (acoustics of sounds, prosody); speech communication (intercultural communication, structure of conversation and temporal characteristics of speech); applied linguistics (language testing). E-mail: hille.pajupuu[at]eki.ee.