Estonian Emotional Speech Corpus: theoretical base and implementation

Conference Paper · January 2012

CITATIONS
9

READS
157

2 authors:

Rene Altrov
Institute of the Estonian Language
23 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Hille Pajupuu
Institute of the Estonian Language
39 PUBLICATIONS **98** CITATIONS
SEE PROFILE

4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS ES³ 2012

Editors

Laurence Devillers	Université Paris-Sorbonne 4, France
Björn Schuller	Technische Universität München, Germany
Anton Batliner	Friedrich-Alexander-University, Germany
Paolo Rosso	Universitat Politècnica de Valencia, Spain
Ellen Douglas-Cowie	Queen's University Belfast, UK
Roddy Cowie	Queen's University Belfast, UK
Catherine Pelachaud	CNRS - LTCI, France

Workshop Organizers/Organizing Committee

Laurence Devillers	Université Paris-Sorbonne 4, France
Björn Schuller	Technische Universität München, Germany
Anton Batliner	Friedrich-Alexander-University, Germany
Paolo Rosso	Universitat Politècnica de Valencia, Spain
Ellen Douglas-Cowie	Queen's University Belfast, UK
Roddy Cowie	Queen's University Belfast, UK
Catherine Pelachaud	CNRS - LTCI, France

Workshop Programme Committee

Vered Aharonson	AFEKA, Israel
Alexandra Balahur	EC’s Joint Research Centre, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Carlos Busso	University of Texas at Dallas, USA
Rafael Calvo	University of Sydney, Australia
Erik Cambria	National University Singapore, Singapore
Antonio Camurri	University of Genova, Italy
Mohamed Chetouani	Université Paris 6, France
Thierry Dutoit	University of Mons, Belgium
Julien Epps	University of New South Wales, Australia
Anna Esposito	IIASS, Italy
Hatice Gunes	Queen Mary University, UK
Catherine Havasi	MIT Media Lab, USA
Bing Liu	University of Illinois at Chicago, USA
Florian Metze	Carnegie Mellon University, USA
Shrikanth Narayanan	University of Southern California, USA
Maja Pantic	Imperial College London, UK
Antonio Reyes	Universidad Politècnica de Valencia, Spain
Fabien Ringeval	Université de Fribourg, Switzerland
Peter Robinson	University of Cambridge, UK
Florian Schiel	LMU, Germany
Jianhua Tao	Chinese Academy of Sciences, China
José A. Troyano	Universidad de Sevilla, Spain
Tony Veale	University College Dublin, Ireland
Alessandro Vinciarelli	University of Glasgow, UK
Haixun Wang	Microsoft Research Asia, China

Estonian Emotional Speech Corpus: theoretical base and implementation

Rene Altrov, Hille Pajupuu
Institute of the Estonian Language
Roosikrantsi 6, 10119 Tallinn, Estonia
E-mail: rene.altrov@eki.ee, hille.pajupuu@eki.ee

Abstract

The establishment of the Estonian Emotional Speech Corpus (EESC) began in 2006 within the framework of the National Programme for Estonian Language Technology at the Institute of the Estonian Language. The corpus contains 1,234 Estonian sentences that express anger, joy and sadness, or are neutral. The sentences come from text passages read out by non-professionals who were not given any explicit indication of the target emotion. It was assumed that the content of the text would elicit an emotion in the reader and that this would be expressed in their voice. This avoids the exaggerations of acted speech. The emotion of each sentence in the corpus was then determined by listening tests. The corpus is publicly available at http://peeter.eki.ee:5000/.
This article gives an overview of the theoretical starting-points of the corpus and their usefulness for its implementation.
Keywords: emotional speech corpus, elicited emotions, non-acted speech, perception of emotions

1. Introduction

The Estonian Emotional Speech Corpus (EESC) is the only publicly available corpus containing samples of Estonian emotional speech. The main purpose of the corpus is to serve research of emotion and language technology applications (see http://peeter.eki.ee:5000/).
The creation of the corpus began by formulating theoretical starting-points (Altrov, 2008), based on overviews of existing emotion corpora and previous emotion research (Campbell, 2000; Cowie & Cornelius, 2003; Douglas-Cowie et al., 2003; Scherer et al., 2001; Ververidis & Kotropoulos, 2006). Several questions concerning the scope of the corpus and data selection had to be answered: 1) Which emotions should the corpus cover? 2) Should the corpus contain spontaneous, elicited, or acted emotions? 3) Should the texts in the corpus be spoken, or read? 4) Which texts should be selected and of what length, content and context? 5) Should the texts be presented by professional, or trained speakers (actors, announcers), or non-professionals (ordinary people)? 6) What size should the corpus be? 7) How many readers/speakers should be used? 8) Whom and how many people should be used as emotion evaluators in the perception tests?

2. Theoretical starting-points and creation of the corpus

The main decisions taken concerning the establishment of the corpus were (Figure 1):
1) Initially three emotions: sadness, anger and joy, plus neutral speech were included in the corpus as being the most useful emotions for language technology applications (Campbell, 2000; Iida et al., 2003). In this corpus these three emotions also include other related similar emotions. Thus, joy includes gratitude, happiness, pleasantness and exhilaration present in the reader's voice; sadness includes loneliness, disconsolation, concern and hopelessness; and anger includes resentment, irony, reluctance, contempt, malice and rage. Neutral speech in the corpus is normal speech without any significant emotion.
2) Simulated emotions and actors were not used due to concerns that actors might overact and use emotions that are too intense and prototypical, and therefore differ from speech that would be produced by a speaker experiencing a genuine emotion (Campbell, 2000; Iida et al., 2003; Scherer, 2003).
Authentic and moderately expressed emotions were to be gathered from text passages read out by non-professionals. The presumption was that the context of the text would stimulate the reader to express the emotion contained therein without them being told which emotion to use (Iida et al., 2003; Navas et al., 2004).
The text passages chosen were journalistic texts, unanimously recognised by readers in a special test, to contain the emotions of joy, anger or sadness. The reason for choosing journalistic texts was that when the corpus was created, it was primarily seen as being a tool for the text-to-speech synthesis of journalistic texts.
The person to read out the texts was chosen very carefully: they had to have good articulation, a pleasant voice and a sense of empathy. Experts were asked to evaluate their articulation. As empathic readers are better at rendering the emotions contained in a text, the candidates were asked to take the empathy test by Baron-Cohen & Wheelwright (2004). Another test was carried out to evaluate the pleasantness of the candidates' voices and listeners were asked to pick the speaker with the most pleasant voice (Altrov & Pajupuu, 2008). Finally, a female voice was chosen and 130 text passages were recorded for the corpus. The passages were segmented into sentences, which were then available to be used in the tests to determine the emotion of sentence.
The emotional sense of each corpus sentence is determined by listening tests. The creators of the corpus were not completely sure how well listeners would do trying to identify the emotions contained in non-acted speech without actually seeing the speaker. Therefore, the participants in the listening tests were carefully chosen to

increase the success rate in the identification of the emotion.

Earlier research implies that more mature listeners may recognise emotions from vocal cues better than younger ones (e.g., students), because emotion recognition is a culture-specific skill that can be acquired only with time (Toivanen et al., 2004). Thus the creators of EESC decided to use Estonians who were over 30 and had spent their lives in Estonia.

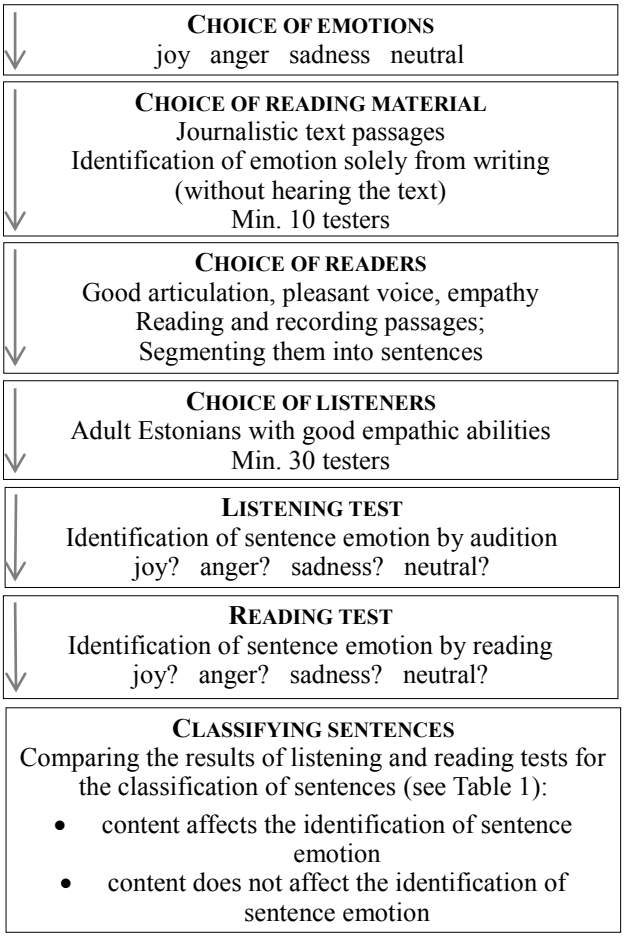

Figure 1: Creation of the EESC.

Previous studies also show that in addition to age, empathy may play a great role in the recognition of emotion (Baron-Cohen & Wheelwright, 2004; Chakrabarti et al., 2006). Relying on the presumption that empathic people are more capable of recognising emotions in voice than non-empathic people (Keen, 2006), candidates were asked to take the empathy test by Baron-Cohen & Wheelwright (2004).

Candidates were also asked, on a voluntary basis, to answer the EPIP-NEO questionnaire (for the Estonian version of the questionnaire see Mõttus et al., 2006) to study links between a person's personality traits and their ability to identify emotions.

The corpus contains 190 registered testers. Collected user data includes: sex, age, education, nationality, mother tongue, language of education, work experience, empathy quotient, and personality profile.

4) The 1,234 sentences in the corpus were used for 14 web-based tests. The underlying principle of the tests was that the content of two successive sentences must not form a logical sequence. Listening test subjects heard isolated sentences without seeing the text and then had to decide which emotion the sentences contained. The available choices were the three emotions: sadness, anger, or joy, or neutral speech.

At least 30 Estonians listened to each sentence.

In 908 sentences more than 50% of listeners identified one and the same emotion, or neutrality.

One issue with the listening tests that needed to be addressed was the role of the content in identifying the emotion of the sentence. Thus, the same sentences were used in 14 reading tests and subjects were asked to decide on the emotion or neutrality of the sentences by reading them (without audio). These subjects were not participants in listening tests.

The emotions identified by the listeners and readers did not always coincide. This led to the establishment of two categories (Table 1):

- sentences where content did not affect emotion identification (the results of reading tests differ from the results of listening tests);
- sentences where content might have affected emotion identification (the results of reading tests coincide with the results of listening tests).

Tests	Joy	Anger	Sadness	Neutral	Not sure	Sentence type in corpus
1. Ehkki Ott minu olemasolust midagi ei teadnud. [Although Ott knew nothing of my existence.]						
By listening	**87.5**	0.0	0.0	12.5	-	**Joy,** no content influence
By reading	4.0	0.0	32.0	32.0	32.0	
2. Ükskõik, mida ma teen, ikka pole ta rahul! [Whatever I do, he is never satisfied!]						
By listening	0.0	14.3	**80.0**	5.7	-	**Sadness**, no content influence
By reading	0.0	**64.3**	35.7	0.0	0.0	
Täiesti mõistetamatu! [Completely incomprehensible!]						
By listening	0.0	**100.0**	0.0	0.0	-	**Anger**, content influence
By reading	0.0	**83.0**	0.0	11.0	5.6	

Table 1: Classification of emotions in the corpus by emotion identification in reading and listening tests (test results in %).

In Table 2 the number of corpus sentences is given by groups.

Emotion	Sentences	Content influence on identification	No content influence on identification
joy	232	163	69
anger	277	177	100
sadness	191	88	103
neutral	208	87	121
unable to identify	326		
Total	1234		

Table 2: Number of sentences in emotion corpus.

Although such double testing of each Corpus sentence is rather time-consuming, it works as a validator for the

corpus. Corpus users can be sure that corpus sentences contain emotions that can be identified during listening. Users can select sentences where emotion is rendered by voice only or sentences where emotion is also rendered by content.

5) The corpus was designed so that it could be used for multiple purposes and extended by adding readers, sentences and emotions.

3. Options for corpus users

Users can search for sentences expressing anger, joy, or sadness, or neutral sentences from the corpus (http://peeter.eki.ee:5000/reports/list/).

Sentences are displayed as text and can be listened to by clicking on them. The identification rate of emotion in each sentence is also displayed.

Queries can be narrowed down to include only sentences in which:

- content did not affect the identification of emotion;
- content might have affected the identification of emotion.

The audio-recordings and text of sentences can be downloaded and saved (wav, textgrid). There are three labelling levels: phonemes, words and pauses, sentences.

4. Implementation details

The corpus is a web-based application that uses freeware: Linux, PostgreSQL, Python, Praat. All data except for audio files have been saved in a PostgreSQL database. The web interface was created and all data processing carried out by using the Python programming language and Pylons web framework. The application can be installed in Windows and Linux systems. The web interface is available for Estonian, English, Finnish, Latvian, Russian and Italian, and can be easily adapted for other languages. For the technical description of the corpus see http://peeter.eki.ee:5000/docs/

5. Preliminary results

Currently the corpus is in a stage where the validity of the theoretical starting-points can be verified and, if necessary, corrections can be made.

1. It has been confirmed that listeners can easily identify moderately expressed emotions from the voice of a non-professional reader. For 73.5% of corpus sentences over 50% of listeners identified one and the same emotion, or decided that the sentence was neutral (Altrov & Pajupuu, forthcoming), see Table 3.

Listening response	Joy	Anger	Sadness	Neutral
Emotional sentences identified by more than 50% of listeners	232	277	191	208
Mean percentage of identification and std	75.4 14.5	73.3 14.6	72.1 14.7	68.3 11.9

Table 3: Statistics of the emotional and neutral sentences identified by the listening test.

2. In the early stages of creating the EESC the decision was made to use people older than 30 as emotion identifiers. This decision relied on the assumption that people who have lived longer in a certain culture are more likely to have acquired the skills of culture-specific expression of emotions. In order to find out if the decision to use older people as corpus testers was justified, Altrov and Pajupuu (2010) compared the results of emotion identification by people older than 30 and younger than 28 and found that the two groups differed significantly. Younger people identified more sentences as neutral. Both groups were also compared with Latvians. The latter identified emotions quite differently from Estonians. From these results it can be said that the identification of emotions really is culture-specific and accurate emotion identification requires spending a longer period in a particular culture. It is therefore wise to use people who have lived in Estonia longer for identifying emotions from vocal expression.

3. Currently a study is being undertaken on how much listeners' empathic abilities affect their ability to identify emotions from vocal expression.

4. Another issue that needs to be addressed is whether classifying corpus sentences according to the influence of sentence content on emotion identification is justified, i.e., if any significant differences can be found between the acoustic parameters of the two groups – "content affects identification" and "content does not affect identification". So far, the corpus material has only been used for studying the difference in intensity of sentence emotions in the two groups. ANOVA analysis has shown that the intensity of sentences expressing anger and joy and neutral sentences in the two groups differ significantly. However, there is no such difference in intensity in sentences expressing sadness (Table 4). Although it is just one acoustic characteristic, it may mean that the content of text affects how an emotion is acoustically expressed, which also means that dividing corpus sentences into two groups is justified.

Pairs: content influences – no content influences	Df	Sum Sq	Mean Sq	F value	Pr(>F)
joy	1	103.00	103.00	5.62	0.0178
Residuals	4189	76707.60	18.31		
anger	1	271.38	271.38	11.92	0.0006
Residuals	5053	114992.73	22.76		
sadness	1	2.80	2.80	0.13	0.7166
Residuals	3467	73757.47	21.27		
neutral	1	591.40	591.40	31.66	0.0000
Residuals	3949	73755.52	18.68		

Table 4: ANOVA results on emotional intensity of sentences in two groups: "content affects identification" and "content does not affect identification".

6. Conclusion

This paper gives an overview of the theoretical base, creation and content of the Estonian Emotional Speech Corpus. The EESC contains 1,234 Estonian sentences that have passed both reading and listening tests. Test takers identified 908 sentences that expressed anger, joy,

sadness, or were neutral. The sentences were divided into two groups: sentences in which content affected the identification of the emotion and sentences in which it did not. Development of the corpus continues. Corpus sentences have also been categorised as positive, negative and neutral. Preparations for extending the corpus by adding video clips with spontaneous speech and their testing are under way. The corpus is freely available and used in the language technological projects for emotional speech synthesis, as well as for recognition of emotions.

7. Acknowledgements

The study was supported by the National Programme for Estonian Language Technology and the project SF0050023s09 "Modelling intermodular phenomena in Estonian".

8. References

Altrov, R. (2008). Eesti emotsionaalse kõne korpus: teoreetilised toetuspunktid. *Keel ja Kirjandus*, 4, pp. 261–271.

Altrov, R., Pajupuu, H. (2008). The Estonian Emotional Speech Corpus: release 1. In F. Cermák, R. Marcinkevicienė, E. Rimkutė & J. Zabarskaitė, *The Third Baltic Conference on Human Language Technologies.* Vilnius: Vytauto Didžiojo Universitetas, Lietuviu kalbos institutas, pp. 9–15.

Altrov, R., Pajupuu, H. (2010). Estonian Emotional Speech Corpus: Culture and age in selecting corpus testers. In I. Skadiņa, A. Vasiļjevs (Eds.), *Human Language Technologies – The Baltic Perspective – Proceedings of the Fourth International Conference Baltic HLT 2010.* Amsterdam: IOS Press, pp. 25–32.

Altrov, R., Pajupuu, H. (forthcoming). Estonian Emotional Speech Corpus: Content and options. In G. Diani, J. Bamford, S. Cavalieri (Eds.). *Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistis.* Amsterdam: John Benjamins.

Baron-Cohen, S., Wheelwright, S. (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,* 34(2), pp. 163–175.

Campbell, N. (2000). Databases of emotional speech. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.), *ISCA Workshop on Speech and Emotions.* Newcastle: North Ireland, pp. 34–38.

Chakrabarti, B., Bullmore, E., Baron-Cohen, S. (2006). Empathizing with basic emotions: common and discrete neural substrates. *Social neuroscience,* 1(3-4), pp. 364–384.

Cowie, R., Cornelius, R.R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication,* 40(1-2), pp. 5–32.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P. (2003). Emotional speech: Towards a new generation of databases, *Speech Communication*, 40, pp. 33–60.

Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication,* 40(1-2), pp. 161–187.

Keen, S. (2006). A theory of narrative empathy. *NARRATIVE*, 14(3), pp. 207–236.

Mõttus, R., Pullmann, H., Allik, J. (2006). Toward more readable Big Five personality inventories. *European Journal of Psychological Assessment,* 22(3), pp. 149–157.

Navas, E., Castelruiz, A., Luengo, I., Sanchez, J., Hernaez, I. (2004). Design and recording an audiovisual database of emotional speech in Basque. In International conference on language resources and evaluation (LREC), Lisbon Portugal, pp. 1387–1390.

Scherer, K.R., Banse, R., Wallbott, H.G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology,* 32(1), pp. 76–92.

Toivanen, J., Väyrynen, E., Seppänen, T. (2004). Automatic discrimination of emotion from spoken Finnish. *Language & Speech,* 47(4), pp. 383–412.

Ververidis, D., Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9), pp. 1162–1181.