Altrov, Rene; Pajupuu, Hille. 2010. "Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers." In *Human Language Technologies – The Baltic Perspective - Proceedings of the Fourth International Conference Baltic HLT 2010*, 25-32, Inguna Skadiņa, Andrejs Vasiļjevs (eds.). Amsterdam: IOS Press.

Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers

Rene ALTROV[1] and Hille PAJUPUU
Institute of the Estonian Language, Tallinn

Abstract. The Estonian Emotional Speech Corpus serves as the acoustic basis for emotional text-to-speech synthesis. Because the Estonian synthesizer is a TTS-synthesizer, we started off by focusing on read texts and the emotions contained in them. The corpus is built on a theoretical model and we are currently at the stage of verifying the components of the model. In the present article we give an overview of the corpus and the principles used in selecting its testers. Some studies show that people who have lived longer in a certain culture can more easily recognize vocal expressions of emotion that are characteristic of the culture without seeing the speaker's facial expressions. We therefore decided not to use people under 30 years of age as testers of emotions in our theoretical model. We used two tests to verify the selection principles for the testers. In the first test, 27 young adults aged under 30 were asked to listen to and identify the emotion (joy, anger, sadness, neutral) of 35 sentences. We then compared the results with those of adults aged over 30. In the second test we asked 32 Latvians listen to the same sentences, and then compared the results with those of Estonians. Our analysis showed that younger and older testers, Estonians and Latvians perceive emotions quite differently. From these test results we can say that the selection principle of corpus testers, using people who are more familiar with Estonian culture, is acceptable.[2]

Keywords. Emotion, vocal expression, ageing, perception of emotions, Estonian

Introduction

Work on the creation of the Estonian Emotional Speech Corpus (EESC) started in 2006 at the Institute of the Estonian Language within the framework of the National Program for Estonian Language Technology. The corpus serves as the acoustic basis for corpus-based emotional text-to-speech synthesis. As the Estonian synthesizer is a TTS synthesizer [1], we focused on read texts and the emotions they contain.

The theoretical model of the EESC relies on research results on emotional corpora and emotions in general [2]. The corpus was created using the following principles:

1. Professional actors were not used because acted emotions are stereotypical and exaggerated and thus different from real-life emotional communication (see [3], [4]). Based on the presumption that listeners can recognize emotions in natural speech quite well, recordings of texts read by ordinary people were used (cf. [5]).

[1] Corresponding Author: Researcher Extraordinary, Institute of the Estonian Language, Roosikrantsi 6, Tallinn, 10119 Estonia; E-mail: rene.altrov@eki.ee.

[2] The study was supported by the National Program for Estonian Language Technology and the project SF0050023s09 "Modeling intermodular phenomena in Estonian".

2. The emotions contained in the corpus sentences were subjected to perception tests.

3. The corpus can be enlarged with readers, sentences and emotions and, in addition to its main function as an acoustic basis for corpus-based emotional speech synthesis, it can be used for other purposes such as research of spoken or written emotions.

4. The corpus is publicly accessible during all its development stages.

1. Corpus Principles

EESC consists of six creation stages [2].

CHOICE OF EMOTIONS. The first step in corpus creation was to choose the emotion categories to be used. We decided to include in the corpus sentences expressing joy, anger and sadness and, to satisfy the needs of speech synthesis, neutral sentences.

CHOICE OF READING MATERIAL. As the main function of the Estonian TTS-synthesizer is to read out journalistic texts, not to have conversations, corpus material came from the Estonian press. Instead of isolated sentences, we decided to record text passages, as the message of a passage facilitates the reader to achieve the right emotional state (see [5]). We chose texts that covered a wide range of topics and tried to avoid colloquial style, keeping in mind that the main function of the synthesizer is to read out written texts.

The next step was to ask a group of people to read the text passages quietly on their own and determine the emotions (joy, anger, sadness) contained in the passages. We included in the corpus passages which contained identifiable emotions.

CHOICE OF READERS. Texts for the EESC were read by a non-actor, a woman with correct pronunciation and a pleasant voice. The pleasantness of her voice was assessed by listeners [6]. In corpus creation we followed the principle that the text itself elicited which emotions were used to read it. We thus did not dictate to the reader which emotions to express but let her decide depending on the text.

CHOICE OF LISTENERS AND LISTENING TEST. Corpus passages were segmented into sentences. As it was up to the reader of the text to choose which emotions to express in each case, each corpus sentence was subjected to a listening test where listeners identified the emotions. A web-based user interface was used for testing. Testers listened to isolated sentences and decided which emotions they contained. They could choose between three main emotions – joy, anger and sadness – and neutrality. Data on testers covered sex, education, nationality, mother tongue, language of education and age.

When the corpus was created, it was not at all certain if it was possible to identify emotions successfully by listening to recorded texts in non-acted Estonian and not seeing facial expressions. To increase the success rate, we chose people older than 30 to be our testers of corpus sentences as research results show that people who have lived longer in a particular culture and have acquired culture-specific skills of expressing emotions are better at vocal emotion recognition (see [7]). That is why we excluded younger people from the group of testers.

In the corpus, sentences are stored together with listening test data. An emotion is considered identified when at least 51% of the participants of the listening test have recognized this emotion in a sentence.

Altrov, Rene; Pajupuu, Hille. 2010. “Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers.” In *Human Language Technologies – The Baltic Perspective - Proceedings of the Fourth International Conference Baltic HLT 2010*, 25-32, Inguna Skadiņa, Andrejs Vasiļjevs (eds.). Amsterdam: IOS Press.

READING TEST. It is likely that vocal emotion recognition is influenced by the semantic context of text as any emotional text contains some emotionally marked words [3]. So far, corpus creators have not given enough credit to content influence on emotion recognition. The issue of content influence is not particularly important when the readers of texts or sentences are told which emotions to express and later the recordings are subjected to listening tests to check if listeners perceive the intended emotions.

However, in the EESC the reader was not asked to express particular emotions. Instead, the listeners were asked to identify the emotion of each sentence. As the corpus mainly serves as an acoustic basis for speech synthesis, it is important to distinguish sentences where the emotions are rendered by vocal expression only. To discover in which cases the emotion is rendered by voice only and in which cases emotion recognition is influenced by text, the same sentences that passed a listening test, were subjected to a reading test with different testers.

The results of the two tests, listening and reading, were compared (see Table 1), and the outcome (if a sentence belongs to the group where emotions were recognized from voice only or to the group where content may have influenced the perception of emotions) was recorded in the corpus database.

Table 1. Principles of emotion classification in the corpus

Tests	**joy**	**anger**	**sadness**	**neutral**	**not sure**[3]	**comment**	**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	-	identified as joy	**Joy,** no content influence
By reading	4.0	0.0	32.0	32.0	32.0	emotion not identified	
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	-	identified as sadness	**Sadness,** no content influence
By reading	0.0	**64.3**	35.7	0.0	0.0	identified as anger	
3. Täiesti mõistetamatu! [Completely incomprehensible!]							
By listening	0.0	**100.0**	0.0	0.0	-	identified as anger	**Anger**, content influence
By reading	0.0	**83.0**	0.0	11.1	5.6	identified as anger	

CONTENT OF CORPUS. The corpus contains sentences expressing joy, anger and sadness, and neutral sentences (currently 579 sentences that have passed both a listening and a reading test). They have been divided into sentences in which the emotion is rendered by voice only and sentences in which emotion recognition may have been influenced by content. The sentences have been segmented and labeled into words and phonemes. The corpus can be found at http://193.40.113.40:5000/ together with a technical description. The user interface is in Estonian, English, Finnish and Latvian.

Currently the EESC is in a state where it is possible to verify the applicability of the decisions made during its creation. We have found proof that emotions can be

[3] “Not sure” was added in the Reading menu for cases where the reader finds it hard to pick an emotion, feeling that the emotion depends more on how the sentence sounds when uttered.

identified in normal, non-acted speech, and we have studied content influence on emotion recognition [8].

In our present study we try to find out if our decision to use corpus testers older than 30 and leave out younger people is well-founded. We look at the relation between emotion recognition and the age and cultural/national background of testers.

2. Age-Related Effects on Emotion Recognition

It is often presumed that as people get older, they become wiser and more experienced through interaction with other people, while their memories get worse and their cognitive abilities slow down. Age also influences a person's ability to understand a communication partner's emotional signals and cues [9]. Age-related studies mainly focus on the recognition of facially expressed emotions and show that as people get older, their ability to identify emotions, especially negative emotions, becomes impaired [9]–[13].

However, there has been little research into the recognition of vocal expression of emotion. Findings indicate that ageing does not necessarily impair the perception of all emotions, and that impairments do not all happen simultaneously [9], [12], [14]. Older people are less accurate at recognizing two negative emotions – anger and sadness [9], [11]. Although there is a small impairment in the perception of positive emotions as well, it only becomes noticeable after the age of 60. As to perceiving neutrality of expression, there are no striking differences between age groups [9].

While it is reasonable to assume that people at the age of 60 and over are likely to have difficulties recognizing different emotions, the latest research results show that it may actually happen earlier. For example, Paumann et al. [12] established remarkable differences in vocal emotion recognition by middle-aged (aged 42–45) and younger (aged 23–25) people. Moreover, Mill et al. [9] discovered that young adults at the ages of 21–30 are slightly less efficient in the identification of sadness and anger than are younger people, but the decline is already significant by the ages of 31–40.

However, not all research results are comparable as it is not always clear how the studies have been conducted and results obtained. There are many unanswered questions such as what reading material was chosen; were professional actors or ordinary people used to express the sentence emotions; were the emotions contained in sentences determined by researchers, for example by asking readers to read neutral sentences with different emotions, or were the readers themselves able to decide on which emotions to use? We believe that these factors have considerable influence on emotion recognition as, for example, it has been determined that the recognition of certain emotions is facilitated by lexical context [12]. It is also easier to identify emotions expressed by actors as these normally sound more stereotypical and thus differ from real emotional expression.

In our study we try to find out if older (aged 30–62) and younger (aged 20–28) people perceive emotions differently when the sentences they hear are read by a non-actor and content influence on emotion recognition has been excluded. We are interested in how justified our decision to use people over 30 as testers of the corpus emotions was. We made this decision assuming that people who have lived longer in a specific culture are better at vocal recognition of culture-specific emotions. We did not consider the possibility that the ability to recognize emotions may fall with age.

Altrov, Rene; Pajupuu, Hille. 2010. “Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers.” In *Human Language Technologies – The Baltic Perspective - Proceedings of the Fourth International Conference Baltic HLT 2010*, 25-32, Inguna Skadiņa, Andrejs Vasiļjevs (eds.). Amsterdam: IOS Press.

3. Material and Method

We composed a listening test of corpus sentences in which testers over 30 had identified joy (10 sentences), anger (10 sentences), sadness (10 sentences) and neutrality (5) and in which content influence on emotion recognition had been excluded (for the principles of sentence selection, see Table 1).

Two groups of testers took the listening test. One group consisted of young Estonian adults (aged 20–28). The other consisted of Latvians. Latvians were used as testers to find out how important it is to live in a specific culture in order to be able to recognize emotions from voice. By comparing the results of the over-30 and under-30 groups and Latvians we can have an insight in how cultural experiences influence emotion recognition. We chose Latvians because Estonian and Latvian cultures share similar values [15] and therefore we can assume that the two nations express emotions similarly and can recognize each other’s emotions from voice alone.

Both groups were asked to listen to isolated sentences without seeing the text and decide which emotion they heard in each sentence. The choices were joy, anger, sadness and neutrality. Listening tests were web-based and were carried out through a user interface for test creation that was connected to the corpus.

The first group comprised 27 testers under 30 years of age with Estonian as their mother tongue (L1). The second group consisted of 32 Latvians.

We compared the results of the two groups with each other and with those of the over-30 Estonians, using R for an Analysis of Variance (ANOVA).

4. Results and Discussion

Corpus sentences were labeled according to the emotion identifications of adult testers (aged over 30). An emotion was considered identified when over 51% of the testers decided in favor of it.

Young adult testers and Latvian testers were also asked to determine the emotion of 35 sentences previously identified by adult testers. (The identification percentage of sentences by each group see http://urve.eki.ee:5000/table1_identification.doc)

To find out if adult testers identify emotions differently from young adults and Latvians, we used ANOVA and logistic regression.

For example, we used the following formula (Eq. (1)) to determine the influence of cultural background on emotion recognition:

```
anova(glm(outcome~nationality, family=binomial, data=t29), test="Chisq"),  (1)
```

where t29 was source data table, and outcome and nationality were columns of the table to which logistic regression was applied. (For source data see http://urve.eki.ee:5000/data.csv)

The results are given in Table 2.

All three groups – adults, young adults and Latvians – differ significantly in identifying emotions. Adults and young adults identify sadness and neutrality in voice similarly. Young adults and Latvians are close in perceiving anger.

Table 2. ANOVA results *(Est – Estonians, LV – Latvians, EstA – adult Estonians, EstY – young adult Estonians)*

Pairs		Df	Deviance	Residuals		P(>|Chi|)
				Df	Deviance	
Est–Lv	**All**	1	99.4299	3095.0000	3899.6600	0.0001
EstA–EstY		1	45.9219	2025.0000	2375.5249	0.0001
EstY–Lv		1	28.4631	2099.0000	2814.6662	0.0001
EstA–Lv		1	145.2425	2064.0000	2517.2849	0.0001
Est–Lv	**Joy**	1	59.9783	894.0000	1147.3442	0.0001
EstA–EstY		1	17.9018	588.0000	712.7969	0.0001
EstY–Lv		1	22.0051	605.0000	818.8808	0.0001
EstA–Lv		1	77.5568	593.0000	727.2071	0.0001
Est–Lv	**Anger**	1	25.4480	875.0000	1159.1030	0.0001
EstA–EstY		1	38.3743	572.0000	701.2394	0.0001
EstY–Lv		1	1.8557	591.0000	820.1342	**0.1731**
EstA–Lv		1	57.4799	585.0000	720.0837	0.0001
Est–Lv	**Sadness**	1	10.7516	881.0000	901.0394	0.0010
EstA–EstY		1	0.5282	575.0000	540.8430	**0.4673**
EstY–Lv		1	6.0636	599.0000	646.3946	0.0138
EstA–Lv		1	9.9839	586.0000	613.7846	0.0016
Est–Lv	**Neutral**	1	14.2618	439.0000	575.5830	0.0002
EstA–EstY		1	3.0013	284.0000	357.7641	**0.0832**
EstY–Lv		1	5.7430	298.0000	406.2836	0.0166
EstA–Lv		1	16.9807	294.0000	381.1156	0.0001

Table 3 presents the confusion pattern in the emotional identification task by cultural background (nationality) and age (mean values in %).

Table 3. Percentage of Estonians and Latvians who identified emotions of sentences according to each emotion target *(Est – Estonians, LV – Latvians, Y – young adult Estonians, A – adult Estonians)*

Nationality	Target	Age	Joy	Anger	Sadness	Neutral
Est	Joy	Y	61.1	7.0	6.0	25.9
		A	77.2	4.8	4.5	13.5
	Anger	Y	1.7	53.4	18.3	26.6
		A	3.2	77.8	4.2	14.8
	Sadness	Y	0.7	9.8	81.0	8.5
		A	2.1	8.9	83.3	5.7
	Neutral	Y	13.8	11.7	11.7	62.8
		A	9.9	9.2	8.5	72.3
Lv	Joy		42.2	19.6	6.5	31.7
	Anger		10.9	47.9	14.2	27.1
	Sadness		2.0	9.8	72.5	15.5
	Neutral		13.5	21.9	15.5	49.0

Altrov, Rene; Pajupuu, Hille. 2010. "Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers." In *Human Language Technologies – The Baltic Perspective - Proceedings of the Fourth International Conference Baltic HLT 2010*, 25-32, Inguna Skadiņa, Andrejs Vasiļjevs (eds.). Amsterdam: IOS Press.

The identification percentage of the target emotion is over 51 for both Estonian adults and young adults, although it is lower for young adults. Young adults perceive considerably more sentences as neutral.

Adults and young adults are closest to each other in identifying sadness (see Table 3). This shows that ageing does not negatively influence the perception of sadness in regular, non-acted speech (cf. [9], [11]).

Latvians identify only sadness in more than 51% of cases. There is a significant difference in how they identify sadness, though (see Table 2). For Latvians, often sentences sound neutral rather than anything else. The fact that Latvians and young adult Estonians identified many sentences as neutral shows that in order to successfully identify an emotion in voice, one has to have a longer experience in how emotions are expressed in a certain culture (cf. [16]).

We asked whether the decision to use testers aged over 30 was right. From our research results, we cannot give a conclusive answer to our research question.

On the one hand, results show that testers aged over 30 and younger testers differ significantly in how they identify emotions. On the other hand neither group has problems with identifying emotions in voice. There was a strong consensus (over 51%, see Table 3) among the two groups about which emotions were heard in sentences, but younger people tend to identify more sentences as neutral. As all sentence emotions contained in the corpus were determined by listeners, we cannot really say which of the two groups was better at emotion recognition. There are two points that support the group of adult testers: 1) the ability of older testers to recognize emotions from voice has not lessened during the process of ageing as they recognize sadness similarly to younger testers (see Table 2); 2) older testers perceive fewer sentences as neutral than do younger testers which shows that older people are better at decoding the culture-specific expression of emotions. The test results for Latvians also confirm the importance of the cultural aspect in emotion recognition as their ability to recognize Estonian emotions from voice was relatively low.

5. Conclusion

In the initial stage of creating the Estonian Emotional Speech Corpus a decision was made to use testers 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. Having studied the relations between testers' age and cultural background and their ability to recognize emotions, and having evaluated the results directly and indirectly we can say that the initial decision is acceptable.

References

[1] M. Mihkla, L. Piits, T. Nurk, I. Kiissel, Development of a unit selection TTS system for Estonian. *Proceedings of the Third Baltic Conference on Human Language Technologies: The Third Baltic Conference on Human Language Technologies.* Vilnius: Vytauto Didžiojo Universitetas. Lietuviu kalbos institutas (2008), 181–187.

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

[3] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: Towards a new generation of databases. *Speech Communication* **40** (2003), 33–60.
[4] K.R. Scherer, Vocal communication of emotion: A review of research paradigms. *Speech Communication* **40**, 1–2 (2003), 227–256.
[5] A. Iida, N. Campbell, F. Higuchi, M. Yasumura, A corpus-based speech synthesis system with emotion. *Speech Communication* **40**, 1–2 (2003), 161–187.
[6] R. Altrov, H. Pajupuu, 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.* Vilnius: Vytauto Didžiojo Universitetas. Lietuviu kalbos institutas (2008), 9–15.
[7] J. Toivanen, E. Väyrynen, T. Seppänen, Automatic discrimination of emotion from spoken Finnish. *Language & Speech* **47**, 4 (2004), 383–412.
[8] R. Altrov, H. Pajupuu, Estonian emotional speech corpus: Content and options. *Rassegna Italiana di Linguistica Applicata*, forthcoming.
[9] A. Mill, J. Allik, A. Realo, R. Valk, Age related differences in emotion recognition ability: a cross-sectional study. *Emotion* **9**, 5 (2009), 619–630.
[10] A.J. Calder, J. Keane, T. Manly, R. Sprengelmeyer, S. Scott, I. Nimmo-Smith, A. W. Young, Facial expression recognition across the adult life span. *Neuropsychologia* **41**, 2 (2003), 195–202.
[11] D.M. Isaacowitz, C.E. Löckenhoff, R.D. Lane, R. Wright, L. Sechrest, R. Riedel, P.T. Costa, Age differences in recognition of emotion in lexical stimuli and facial expressions. *Psychology and Aging* **22**, 1 (2007), 147–159.
[12] S. Paulmann, M.D. Pell, S.A. Kotz, How aging affects the recognition of emotional speech. *Brain and Language* **104**, 3 (2008), 262–269.
[13] L.H. Phillips, R.D.J. MacLean, R. Allen, Age and the understanding of emotions: Neuropsychological and sociocognitive perspectives. *Journal of Gerontology B: Psychological Sciences and Social Sciences* **57**, 6 (2002), 526–530.
[14] P. Laukka, P.N. Juslin, Similar patterns of age-related differences in emotion recognition from speech and music. *Motivation and Emotion* **31** (2007), 182–191.
[15] M. Huettigen, Cultural dimensions in business life: Hofstede's indices for Latvia and Lithuania. *Baltic Journal of Management* **3**, 3 (2008), 359–376.
[16] M.D. Pell, S. Paulmann, C. Dara, A. Alessari, S.A. Kotz, Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics* **37** (2009), 417–435.