Development and Evaluation of the Emotional Slovenian Speech Database - EmoLUKS

Tadej Justin[1], Vitomir Štruc[1], Janez Žibert[2], and France Mihelič[1(✉)]

[1] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
{tadej.justin,vitomir.struc,france.mihelic}@fe.uni-lj.si
[2] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, 6000 Koper, Slovenia
janez.zibert@upr.si

Abstract. This paper describes a speech database built from 17 Slovenian radio dramas. The dramas were obtained from the national radio-and-television station (RTV Slovenia) and were given at the universities disposal with an academic license for processing and annotating the audio material. The utterances of one male and one female speaker were transcribed, segmented and then annotated with emotional states of the speakers. The annotation of the emotional states was conducted in two stages with our own web-based application for crowd sourcing. The final (emotional) speech database consists of 1385 recordings of one male (975 recordings) and one female (410 recordings) speaker and contains labeled emotional speech with a total duration of around 1 hour and 15 minutes. The paper presents the two-stage annotation process used to label the data and demonstrates the usefulness of the employed annotation methodology. Baseline emotion recognition experiments are also presented. The reported results are presented with the un-weighted as well as weighted average recalls and precisions for 2-class and 7-class recognition experiments.

Keywords: Emotional speech database · Emotion recognition · Database development

1 Introduction

Paralinguistic information is commonly defined as speaker- or speech-related information that cannot be described properly with phonetic or linguistic labels. Examples of paralinguistic information include the level of intoxication of a speaker, the speaker's mood, his/her interests, or the emotional state of the speaker among others.

The design and development of speech databases, which include paralinguistics labels, demands interdisciplinary cooperation [1], [2], [3]. One of the most demanding parts during the design stage is to properly define the paralinguistic labels. For example, when the selected paralinguistic labels are the emotional

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 351–359, 2015.
DOI: 10.1007/978-3-319-24033-6_40

states of the speakers (such as in our case), discrete states need to be defined that can be associated with the paralinguistic labels. However, this represents a difficult task, since there is no generally established definition of what an emotional state is. In such cases, developers commonly resort to examples of good-practice from the literature, e.g., [4], [5],[6] which define examples of emotional states, or rely on expert knowledge and guidelines for building emotional databases, e.g.,[7].

The development of an emotional database is, in general, guided by the research goals. Clearly, different datasets are needed when studying human emotions from a theoretical perspective (where the goal is to understand how emotions are related to the psychological or biological processes in humans) or when developing applications that try to take the expressed emotions into account during their operation. The latter is also the case with speech technologies, such as emotional speech synthesis or speech recognition from emotionally colored speech, where the goal is to either make the synthesized speech sound more natural or improve the performance of speech recognition systems by accounting for the variability in the speech signal induced by changes in the emotional state of the speaker. With speech technologies, emotional databases typically consist of recordings and transcriptions of speech with additional paralinguistic labels [8].

Over the last few years the field of speech technologies has seen increased interest in modeling techniques and approaches that make use of paralinguistic speaker information [9]. Increased popularity in applications capable of natural human-computer interaction (HCI) using speech can also be observed over the last decade. Since speech technologies are for the most part language dependent, it is important to have suitable resources at ones disposal. Building a speech database with paralinguistic labels (in the form of annotated emotional states of the speakers) is the first steps when trying to improve the naturalness of the existing Slovenian speech synthesis systems or to develop emotion recognition systems for Slovenian speakers.

In this paper we focus on the development of an emotional speech database for the Slovenian language for applicative use - primarily in emotionally colored speech synthesis. We evaluate the annotated emotional speech material and present a use-case for the database, i.e., automatic emotion recognition from speech. We also elaborate on the importance of the two-stage annotation process of the emotional utterances and finally present the Slovenian emotional speech database - EmoLUKS.

The rest of the paper is structured as follows: in Section 2 we describe the preparation and annotation of the database and present a brief analysis of the annotated speech material. We evaluate the annotation of the EmoLUKS database and present emotion recognition experiments in Section 3. Finally, we conclude the paper in Section 4 with some final comments and directions for future work.

2 Emotional Speech Database - EmoLUKS

The development of emotional speech databases typically follows one of two different approaches in terms of design: *i)* With the first approach utterances

of professional actors, who are capable of correctly articulating natural speech and imitate different emotional states, are read and then recorded. In general, such databases are typically comprised of predefined sets of sentences, which are commonly extracted from a large, language-dependent text corpus and selected in a way that balances the distribution of the base phonetic units in one of the target languages. The emotional labels of the sentences are commonly defined before the recording stage. Such an approach results in databases of simulated or acted emotions. *ii)* The second approach relies on the transcription, segmentation and annotation of pre-recorded speech material. In case the pre-recorded speech material corresponds to natural speech, the annotated database represent a database of spontaneous emotional states of speakers, while it may again represent a database of acted emotions if the pre-recorded data is read, such as in the case of radio shows or dramas. The main difference between the first and second approaches is in the way the emotional states of the speakers are annotated. The first one has predefined emotional labels, which are commonly evaluated with perception tests. With the second approach, annotators are commonly employed to annotate the utterances (again with perception tests), but the final labels of the speech utterances are decided based on a majority vote.

In our case we adopted the second approach, and used radio dramas to build a database of acted emotional speech. We selected sentences as the basic annotation units and decided on the final emotion labels based on a majority vote over the labels assigned to the sentences by the annotators.

2.1 Database Description and Preparation

With the help of the national radio-and-television station of Slovenia, i.e., RTV Slovenia, we obtained radio-drama recordings that were produced in a professional studio. We manually transcribed and segmented the radio dramas and also annotated the non-lexical data, which are commonly used in radio dramas, such as background music, various background noises, and various added audio effects. Additionally, we took extra caution when marking and segmenting the speaker's non-lexical sounds, such as crying, breathing, laughter, etc. For this process, the Transcriber annotation tool [10] was used. Once we obtained the transcribed and segmented audio, we extracted the utterances of each speaker based on full-sentence units that included only clear speech. In this way we obtained speech segments that are not too long and with enough contexts for annotating the emotional state of the speakers. Such utterances were prepared for further processing and annotation of the speaker's emotional state.

We transcribed 17 radio dramas with an approximate total time of 12 hours and 50 minutes. The transcribed material includes the segmentation of 16 speaker identities (5 female and 11 male) that produced at least 3 minutes of clear speech. From the transcribed and segmented material we extracted the utterances of one male (01m_av) and one female speaker (01f_lb) (clear speech only) for a total duration of 62 minutes for the male and 15 minutes the female speaker.

2.2 Defining the Emotional States

Each research field that interacts or deals with human emotions needs a proper definition of what an emotional state is. The differences in the theoretical models, on which the theory of emotions is based [11], clearly show how emotions can be subject to different interpretations. In the literature we found four different perspectives on emotional states [12]. The use of each perspective also determines the different relations between the emotional categories. The main assumption for modeling the delimitations of the emotional categories is that the differences between the observed emotional experiences of one category are smaller than the differences between the emotions from different categories. Various aspects to modeling the emotional relations are presented in [11].

We focus on the discretization of the emotional states based on Darwin's perspective [11]. Hence, we assume that there are some basic emotional states from which the basic discrete models of emotional categories were developed. Such a representation is one of the most popular approaches for presenting the emotional space. In this way we determined the discrete basic emotions in different emotional categories for annotating the recordings of radio drama: sadness, joy, gust, anger, fear, surprise and neutral.

2.3 Database Annotation Through Crowd-Sourcing

Annotating the emotional states from speech can be achieved with expert knowledge. While an expert on emotional states can annotate each speech signal, more objective labels can be obtained if several experts provide their expertise and annotate the data. Since there is no generally accepted definition of emotional states and due to the fact that all people are able to interpret human emotion to at least some degree, there is no need to look for experts in the field. Instead ordinary people can be asked to label the data and decide which emotions are expressed in each of the speech utterances.

The annotating procedure for speech material is a relatively time-consuming process. Therefore, it is advisable to provide a software application that allows to annotate utterances in a fast and reliable way. It is also desirable to allow the annotators to pick the annotating time by themselves. In recent years the most attractive way to approach such problems is to allow multiple annotators to annotate from any location thought the web. Such an approach is called crowd-sourcing [13] and has became the most popular solution for annotating different databases in the past decade, especially those meant for further processing or data mining.

Inspecting the literature and the available crowd-sourcing applications, we decided that none of these applications suited our needs - to allow volunteers to perform fast and easy annotation of the speech database. Therefore, we decided to develop our own crowd-sourcing application for audio or video resources.

Our web-based crowd-sourcing application was developed with the help of the available open-source software. It was designed as an add-on for the well-established Content Management System (CMS) Plone (er. 4.3.3) [1] based on

[1] http://plone.org

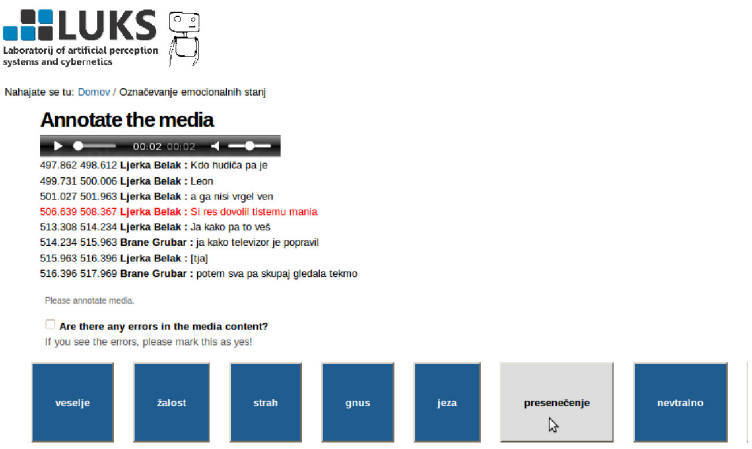

Fig. 1. A screen shoot of an annotator submitting decision "surprise" for the current speech utterance. An annotator is able to read the full context dialog, where the transcription of the current utterance is colored red. The annotator can also mark if the utterance contains an error in the transcription or segmentation.

the Zope [2] application server. The developed crowd-sourcing application consists of two different types of access. The first is designed for editors/administrators who can prepare the annotation procedure through the web. The second tries to offer the annotator a user-friendly annotation experience. Anonymous annotating is not permitted due to academic license associated with the database. The annotators can annotate the resource over a longer time period by simply logging-in/logging-out and starting at the last annotated audio or video file. The demonstration of the application for crowd-sourcing the audio or video signals developed at LUKS is available at http://emo.luks.fe.uni-lj.si. A snapshot of the user interface of the application is shown in Fig. 1.

The developed application works as follows. The speech utterance is randomly picked from the predefined media files that need to be annotated. It is automatically played when the submission form is loaded and the annotator can easily pass his/her decision with a simple mouse click. When the annotator makes the decision, the next media for the annotation is loaded.

We asked 5 volunteers (2 female and 3 male) to annotate the speech material. The annotation labels were picked as discrete basic emotions with an added normal state. During the process of annotation, all the evaluators wore headphones.

2.4 Analysis of the Annotated EmoLUKS Database

The development stage for the EmoLUKS emotional database was already described in [14]. Here, we try to present the difficulties of the first-stage

[2] http://zope.org

annotation procedure, point to some possible improvements and suggest further steps to obtain better and more accurate labels for the speech material.

After the first stage of the annotation process, it was not possible to decide on the paralinguistic label for 18.47% of the utterances using our majority-voting strategy, since no emotional state could be selected as a winner after the vote. To mitigate this issue, we decided to ask the same volunteers to annotate the problematic utterances again in the second stage of the annotation process.

The reported result in this paper compare the first and the second-stage annotation of the emotional states of the speakers and consolidate the labels for the EmoLUKS database. Each stage of the two-stage annotation process was conducted by the same annotators, but in different time periods. The first stage included 1010 utterances from one male and one female speaker from the segmented and transcribed radio-drama material. The second stage included 988 utterances from the same speakers as in the first stage. While annotating the first stage, the annotators reported some errors in the transcription and/or segmentation. All reported errors were corrected and the files were once again annotated in the second stage. Also, some utterances, which were too long, but were part of the first annotation stage, were segmented again from full sentences into shorter, but still meaningful parts.

Overall, the second stage included 421 of the corrected utterances from the first stage due to segmentation and/or transcriptions errors. The other 542 utterances were the same as in the first stage of the annotation process. In the second annotation stage we also included 25 utterances that were not included in the first stage. As can be seen from Table 1, where a brief summary of the annotation stages is presented, both annotation stages resulted in approximately the same fraction of utterances, where no winning label could be selected after the majority vote and, hence, the utterances had to be assigned to the "Undecided" category (18.47% vs. 19.53%). Since many of the utterances annotated in the second stage overlapped with the (problematic) utterances that were already annotated in the first stage, ten labels were available for these utterances to decide on a winning label. Clearly, this fact resulted in a reduction of the percentage of utterances in the "Undecided" category and consequently improved the percentage of utterances in all other categories (except for Disgust).

Other information that can be seen from Table 1 is: *i)* the number of utterances per speaker and annotation stage (marked as Utt. #), *ii)* the total duration of the annotated speech material per speaker and annotation stage (marked as Dur.), and *iii)* the fraction of annotated speech material (in terms of duration) per emotional category/label for each speaker and each annotation stage.

3 Baseline Emotion Recognition Experiments

In this section we present one possible use-case for the EmoLUKS database, i.e., developing and testing emotion recognition systems for affective computing

Table 1. Overview of the annotated material in the two-stage annotation process for the Slovenian emotional speech database EmoLUKS based on the annotators majority vote. Here the abbreviations Surp., Disg., Ang.. Neut., Sad., Undec. stand for Surprise, Disgust, Anger, Neutral, Sadness, and Undecided, respectively.

Eval.	Spk.	Utt. #	Dur.	Fraction of annotated speech material [% od total duration]							
				Joy	Surp.	Disg.	Ang.	Neut.	Fear	Sad.	Undec.
1.stage	01m_av	762	1:01:29	8.53	11.02	1.18	14.44	36.48	5.38	4.86	18.11
	01f_lj	348	14:56	11.78	14.94	4.02	28.45	11.21	8.05	2.30	19.25
	sum	**1110**	**1:16:24**	**9.55**	**12.25**	**2.07**	**18.83**	**28.56**	**6.22**	**4.05**	**18.47**
2.stage	01m_av	733	49:12	8.59	9.69	1.36	12.55	35.06	9.41	3.14	20.19
	01f_lj	255	8:27	10.59	17.25	6.27	21.96	12.94	10.98	2.35	17.65
	sum	**988**	**57:39**	**9.11**	**11.64**	**2.63**	**14.98**	**29.35**	**9.82**	**2.94**	**19.53**
EmoLUKS	01m_av	975	1:02:31	8.72	11.59	1.03	15.69	39.69	7.79	4.10	11.38
	01f_lj	410	14:44	12.44	16.83	4.15	27.56	13.17	10.98	3.90	10.98
	sum	**1385**	**1:17:16**	**9.82**	**13.14**	**1.95**	**19.21**	**31.84**	**8.74**	**4.04**	**11.26**

Table 2. Speaker-dependent emotion-recognition results performed on the EmoLUKS speech material. UA (un-weighted average) and WA (weighted average). The results are presented in the form of the mean values and standard deviation of the performance metrics over all five folds.

Speaker label	Problem	Recall		Precision	
		UA	WA	UA	WA
01m_av	**2 class**	0.70 ± 0.04	0.70 ± 0.04	0.70 ± 0.04	0.71 ± 0.04
	7 class	0.27 ± 0.04	0.50 ± 0.05	0.30 ± 0.06	0.47 ± 0.05
01f_lb	**2 class**	0.54 ± 0.04	0.79 ± 0.03	0.56 ± 0.05	0.77 ± 0.02
	7 class	0.32 ± 0.04	0.44 ± 0.05	0.31 ± 0.04	0.42 ± 0.05

applications. Towards this end, we use the OpenSMILE feature extractor [15] and its predefined configuration to extract the same features as were used in the baseline experiments during the Interspeech 2009 Emotion Challenge [16]. The idea behind this experiments is to use a state-of-the-art emotion recognition approach to demonstrate the difficulty of the database for the problem of emotion recognition.

Since the obtained radio-drama recordings were professionally recorded, but during different time periods with different producers and directors, we first normalized all the annotated utterances. The normalization and equalization of the gain was conducted with SoX (Sound eXchange)[3], using its default values. We used a 5-fold stratified cross-validation scheme to evaluate the SVM classifier [17] integrated into the WEKA Data Mining Toolkit [18], with default parameters. Since currently only speech material from one male and one female speaker is annotated, we performed speaker-dependent experiments. Thus, separate SVM classifiers ware trained for each speaker. The classification results are presented in Table 2. They represent two-class classification results for the neutral and the non-neutral emotional states and the seven-class emotional state classification

[3] http://sox.sourceforge.net

problem, where the reference label represents the majority vote of the annotated speech utterances.

Note that the performance of the recognition experiments is quite low. These results may partially be ascribed to the fact that for some emotional classes (such as "Disgust") only a small amount of data is available in the current version of the database, but also show how challenging this database is.

4 Conclusion and Feature Work

In this paper we presented our current efforts towards building a speech database with paralinguistic information. Specific attention was given to the two-stage annotation process for the emotional utterances from one male and one female speaker. Our analysis suggests that the efforts involved in setting up a second annotation stage were repaid. With the consolidated labels from the first- and second-stage of the annotation we were able to significantly decrease the amount of utterances that could not be labeled with a majority vote (undecided category). As indicated in Subsection 2.1, we still have some speech material from different speakers available for the annotation of the emotional states of the speaker. Therefore, we plan to expand the current EmoLUKS database as part of our future work.

References

1. Gajšek, R., Štruc, V., Mihelič, F., Podlesek, A., Komidar, L., Sočan, G., Bajec, B.: Multi-modal emotional database: AvID. Informatica (Slovenia) **33**(1), 101–106 (2009)
2. Batliner, A., Biersack, S., Steidl, S.: The prosody of pet robot directed speech: evidence from children. In: Proc. of Speech Prosody, pp. 1–4 (2006)
3. Koolagudi, S., Rao, K.: Emotion recognition from speech: a review. International Journal of Speech Technology **15**(2), 99–117 (2012)
4. Gajšek, R., Štruc, V., Dobrišek, S., Mihelič, F.: Emotion recognition using linear transformations in combination with video. In: 10th INTERSPEECH (2009)
5. Gajšek, R., Žibert, J., Justin, T., Štruc, V., Vesnicer, B., Mihelič, F.: Gender and affect recognition based on GMM and GMM-UBM modeling with relevance MAP estimation. In: 11th INTERSPEECH (2010)
6. Dobrišek, S., Gajšek, R., Mihelic, F., Pavešič, N., Štruc, V.: Towards efficient multi-modal emotion recognition. International Journal of Advanced Robotic Systems **10**(53), 1–10 (2013)
7. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. Speech Communication **40**(1–2), 5–32 (2003)
8. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: Paralinguistics in speech and language – state-of-the-art and the challenge. Computer Speech & Language **27**(1), 4–39 (2013)
9. Yamashita, Y.: A review of paralinguistic information processing for natural speech communication. Acoustical Science and Technology **34**(2), 73–79 (2013)

10. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication **33**(1–2), 5–22 (2001)
11. Cornelius, R.R.: The science of emotion: Research and tradition in the psychology of emotions. Prentice-Hall, Inc. (1996)
12. Cornelius, R.R.: Theoretical approaches to emotion. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion (2000)
13. Howe, J.: The rise of crowdsourcing. Wired Magazine **14**(6), 1–4 (2006)
14. Justin, T., Mihelic, F., Žibert, J.: Development of emotional Slovenian speech database based on radio drama – EmoLUKS. In: Language Technologies. Proceedings of the 17th International Multiconference INFORMATION SOCIETY - IS 2014, vol. G, Institut "Jožef Stefan" Ljubljana, pp. 157–162 (2014)
15. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 835–838. ACM (2013)
16. Schuller, B., Steidl, S., Batliner, A.: The Interspeech 2009 emotion challenge. In: 10th INTERSPEECH, pp. 312–315 (2009)
17. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation **13**(3), 637–649 (2001)
18. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)