Emotion Recognition in Real-World Support Call Center Data for Latvian Language

Eduards Blumentals[1], Askars Salimbajevs[1,2]

[1] *Tilde SIA, Vienibas gatve 75a, Riga, Latvia*
[2] *Faculty of Computing, University of Latvia, Raina bulvaris 19, Riga, Latvia*

Abstract

Emotion recognition from speech is a research area that focuses on grasping genuine feelings from audio data. It makes it possible to extract various useful data points from sound that are further used to improve decision-making. This research was conducted to test an emotion recognition toolkit on real-world recording of phone calls in Latvian language. This scenario presents at least two significant challenges: mismatch between real-world data and "artificially" created data, and lack of the training data for the Latvian language. The study mainly focuses on investigating training data requirements for successful emotion recognition.

Keywords

datasets, neural networks, speech, emotion recognition

1. Introduction

Nowadays, emotion recognition from speech is a highly relevant topic. It is used in a wide variety of applications from businesses to governmental bodies. For example, in call centers it helps to monitor client support quality and to study clients' reaction to certain emotional triggers. Multiple studies have been conducted on emotion recognition from speech signal. However, most of the papers investigate machine learning model performance on public artificially created datasets such as EMODB[1], IEMOCAP[2], TESS[3], and RAVDESS[4]. Although this approach ensures a common benchmark, it ignores the fact that in the real world speech data is not as clear or well-defined. A few papers such as Kostulas et al.[5], Dhall et al.[6] and Tawari et al.[7] aim to address this issue.

When it comes to emotion recognition from speech in Latvian language, the literature is even shallower. Neither datasets nor well-recognized research on the topic exists for Latvian language. Therefore, this study investigates how a deep learning emotion recognition model performs on a Latvian language dataset comprised of real phone calls. The goal of this research is to see whether typical deep learning architecture can handle the task of detecting emotions from real speech. The paper evaluates model performance with different training data setups. In addition it evaluates human error to estimate the difficulty of the exercise for an untrained person.

2. Data

The main dataset used in this paper consists of technical support phone calls. Recordings are done with an 8 kHz sampling rate and a single channel. The dataset included audio recordings of 39 conversations, that held in Latvian which were further separated into 6,171 segments. Qualitative analysis of telephone conversations was performed, annotating in several layers' potential affective features - affect dimensions, linguistic units, paralinguistic units etc. A total of 11 synchronous annotation layers were created for each segment.

Each segment had two parameters valence and activation, where valence measures how positive or negative the emotion is, and activation measures its magnitude. These parameters were assigned by trained individuals (pedagogy and psychology students and professors). Detailed description of dataset creation process and qualitative analysis of the dataset is presented in [8].

Based on the given dimensions segments were assigned to nine categories: happy, surprised, angry, disappointed, sad, bored, calm, satisfied and neutral following the approach proposed by Russell and Barrett[9]. Therefore, the problem is transformed from regression into a multiclass classification.

An insignificant number of observations in several emotion categories necessitated proceeding with the five most represented emotions. Table 1 summarizes the final dataset used in this paper. This dataset is further divided into train (80%) and test (20%) sets.

Additionaly, several public emotional speech datasets were included in the research. EMODB, IEMOCAP, TESS, and RAVDESS were used to increase the dataset size, as well as to see how our model performs compared to the state-of-the-art models.

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland

eduards.blumentals@tilde.lv (E. Blumentals);
askars.salimbajevs@tilde.lv (A. Salimbajevs)
0000-0003-0165-0868 (A. Salimbajevs)

Table 1
Data Summary

Emotion	Observation count
Surprised	2281
Angry	2208
Neutral	591
Happy	359
Sad	105

3. Methodology

This paper investigates a deep learning approach to emotion recognition. Each input audio was converted into the 39-dimensional mel-frequency cepstral coefficients (MFCC) feature vector and passed through the model. Due to the relatively small dataset size, a shallow neural network was used to prevent overfitting. The final model was comprised of two LSTM layers, two fully connected layers and a softmax output layer. For additional regularization, a 30% dropout after each layer was added. Figure 1 displays the model architecture.

Categorical cross-entropy was used as a loss function and an Adam optimizer was used for backward propagation. The model training was performed using batches of 64 observations. Each model was trained for 200 epochs with validation after each epoch. Next model weights that yielded the highest accuracy were retrieved. Finally, all trained models are compared based upon their accuracy on the test set.

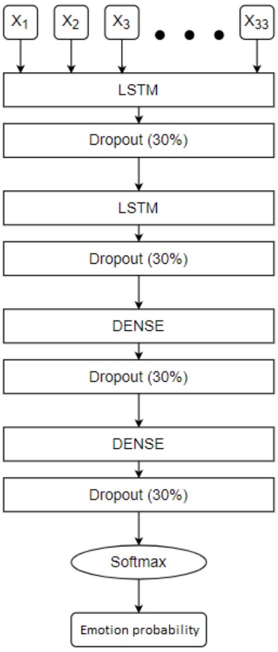

Figure 1: Model Architecture

4. Results

4.1. Validating the Model

First, model performance on public datasets was evaluated. TESS and RAVDESS were combined into one dataset, separated into train (80%) and test (20%) sets and used to train the model for 200 epochs. The final test accuracy (Figure 2) was 86.02%. In addition, a similar experiment was conducted with an IEMOCAP dataset which was comprised of conversations between actors. The final test accuracy (Figure 3) was 60.65% which is slightly lower than state-of-the-art[10]. From this, one can conclude that the deep learning architecture used in this paper performs reasonably well on public "Wizard of Oz" datasets.

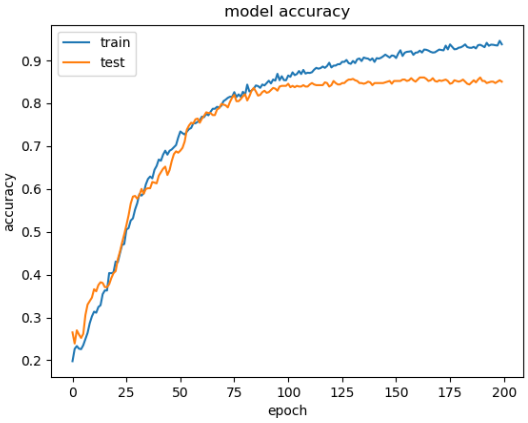

Figure 2: Model Accuracy on TESS & RAVDESS Test Set

4.2. Model Accuracy on Real-World Latvian Data

Next, the impact of changes in training data volume on test accuracy was evaluated to see whether the performance of models can be improved by simply supplying additional data. In this experiment, the model was trained and evaluated on real world audio recordings from a Latvian support call center. The model was trained on different portions of the train set collected from phone call data. It started at 50% volume and moved towards the full train set with a step of 10 percentage points. The results of this experiment are displayed in Figure 4. Seemingly, in the case of this research, increasing data volume would not yield significantly better results.

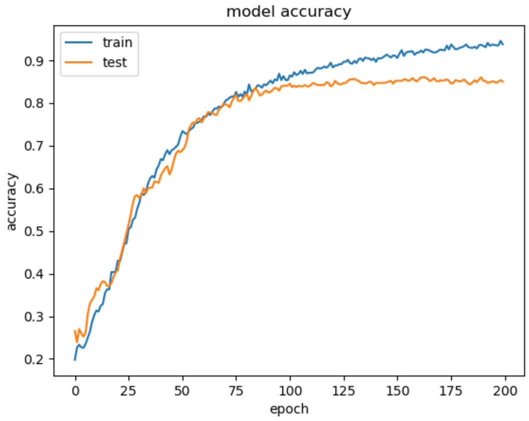

Figure 3: Model Accuracy on IEMOCAP Test Set

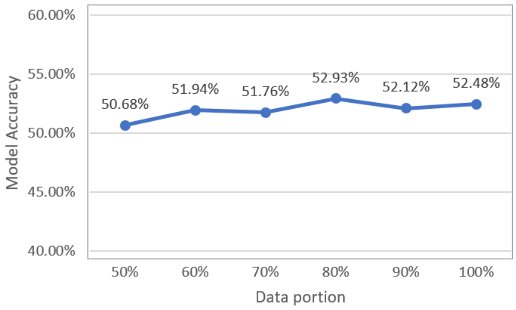

Figure 4: Model Accuracy Depending on Training Data Volume

4.3. Using Additional Training Data Sets

The main goal of the following experiments was to understand if adding additional data from public datasets can improve the performance of models. Because phone call data is recorded with 8 kHz sampling rate, but public datasets are 16 kHz, following 4 experiments were performed.

In the first experiment, the model was trained on the phone call data in its original format. In the second experiment, IEMOCAP, TESS, RAVDESS and EMODB were downsampled to 8 kHz and added to the train set. In the third experiment, phone call data (both train and test) were upsampled to 16 kHz. In the fourth experiment, IEMOCAP, TESS, RAVDESS and EMODB were added to the upsampled train set. The results of those experiments are summarized in Table 2.

Table 2
Training with Additional Data

Training Data	Sampling Rate	Accuracy
Phone calls only	8 kHz	53.83%
All data	8 kHz	50.68%
Phone calls only	16 kHz	53.47%
All data	16 kHz	50.41%

4.4. Obtaining Human Error

Finally, an untrained human person was asked to guess the emotions in the same test set to estimate the human error. Audio segments were presented in random order, so that the person can not analyse the overall semantics and context of the conversations, and have to rely solely on the acoustics, similarly to the deep learning model. The test accuracy ended up being 22.72% which indicates that predicting emotions in the random segments of phone calls is not a trivial exercise even for a human.

5. Conclusions

This research investigated emotion recognition from real world phone call data. The output of the research can be summarized according to the following points:

- Emotion recognition on real-world data is a more difficult exercise than emotion recognition on artificially created datasets
- The model architecture proposed in this paper is capable of surpassing a untrained human-level error for the given exercise
- Augmenting training phone call data with artificially created datasets does not seem to help to improve model performance
- At this stage increasing data volume twofold marginally improves model performance
- Upsampling and downsampling the audio data neither improves nor worsens the performance of the models

For further research it might be worth trying to increase the training dataset further (at least 500-1000% increase). Given the untrained human-level error, it seems that in order to predict emotions accurately, even human needs more context than a single utterance, preferably whole conversations. Therefore, increasing input context is interesting avenue for the follow-up work. Furthermore, defining an emotion as a set of dimensions and predicting each dimension separately might improve forecasting accuracy.

Acknowledgments

The research leading to these results has received funding from the research project "Competence Centre of Information and Communication Technologies" of EU Structural funds, contract No. 1.2.1.1/18/A/003 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 2.9. “Automated multilingual subtitling”.

References

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, A database of german emotional speech, in: Ninth European Conference on Speech Communication and Technology, 2005.

[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359.

[3] M. K. Pichora-Fuller, K. Dupuis, Toronto emotional speech set (TESS) (2020). URL: https://doi.org/10.5683/SP2/E8H2MF. doi:10.5683/SP2/E8H2MF.

[4] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (2018) e0196391.

[5] T. Kostoulas, T. Ganchev, N. Fakotakis, Study on speaker-independent emotion recognition from speech on real-world data, in: Verbal and nonverbal features of human-human and human-machine interaction, Springer, 2008, pp. 235–242.

[6] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: Proceedings of the 15th ACM on International conference on multimodal interaction, 2013, pp. 509–516.

[7] A. Tawari, M. M. Trivedi, Speech emotion analysis in noisy real-world environment, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 4605–4608.

[8] E. Vanags, A qualitative analysis of affect signs in telecommunication dialogues, in: The 79th International Scientific Conference of the UL section Psychological well-being, 2021.

[9] J. A. Russell, L. F. Barrett, Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant., Journal of personality and social psychology 76 (1999) 805.

[10] Y. Wang, J. Zhang, J. Ma, S. Wang, J. Xiao, Contextualized emotion recognition in conversation as sequence tagging, in: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2020, pp. 186–195.