Emotion Recognition Using Audio Speech Signal

Maj Smerkol
Jožef Stefan Institute
Jadranska cesta 39
Ljubljana, Slovenia
maj.smerkol@ijs.si

Mitja Luštrek
Jožef Stefan Institute
Jadranska cesta 39
Ljubljana, Slovenia
mitja.lustrek@ijs.si

ABSTRACT
Emotion recognition is an important part of affective-aware applications. Specifically, using audio speech signal has the advantage of being compatible with applications using a natural language interface. There are multiple valid representations of emotions. We propose a new representation aimed at using differently labeled databases jointly. We include a short overview of some of the available databases and methods for feature extraction and selection. Both classification of emotions and regression in 2D emotional space are discussed. We concentrate on using neural networks for both tasks. Regression provides good results but is hard to interpret while classification is more robust.

Categories and Subject Descriptors
I.5 [**Pattern recognition**]: Neural nets; I.5.2 [**Design methodology**]: Classifier design and evaluation

Keywords
Emotion recognition, Neural networks, Affective computing

1. INTRODUCTION
Nowadays applications such as personal digital assistants are becoming more popular. Some also utilize natural language interfaces. Next step in this direction seems to be affective computing - applications that can detect human emotions. Such applications can enrich the user experience by responding according to the user's current mood and perhaps even detect when the user is not happy with the application's functioning. However, in order to implement such applications we need to first be able to understand the user's emotions. This, in conjunction with other knowledge (such as user's daily routines and other contextual information) makes it possible to detect certain mental health problems, such as depression or bipolar disorder, shown by Osmani et al. [8].

Models we are developing will be used in an emotionally-aware virtual assistant application. Our priority is to deliver information that can be acted upon in order to better the user experience. Application's target population are people from Italy, Spain and Denmark.

1.1 Representations of Emotions
When talking about emotions in the context of affective computing, we first need to consider how to represent human emotions. In psychology, there exist many different theories about human emotions. We can choose a discreet representation of emotions or a contiuous representation in some space of emotions. In first case, we define different categories that represent different emotional states. The most widely known categorization is Paul Ekman's basic emotions. Ekman studied facial expressions of emotions across different cultures and came to the conclusion that there are six basic emotions that are expressed equally across cultures. Those are sadness, happiness, anger, fear, disgust and surprise.

On the basis of Ekman's work others proposed different models. Some of them have a different set of categories. Others use a continuous representation in two, three or four dimensional spaces. There is Plutchik's wheel of emotions that represents emotions as four pairs of exclusive categories, that are treated as four axes along which emotions are spread. Emotions are represented as points in this space. J. Russel proposed a different model, a two dimensional space. Dimensions are arousal, which represents how active one feels, and valence, representing pleasurableness of the emotion.

For our purpose, we prefer classification robustness over precision. We don't need very fine-grained information to better the user experience of the application. Therefore in order to use as much training data as possible, we propose a four-class representation of emotions. The main idea of this representation is to be able to easily transform labels in other representations into a common one. Classes correspond to quadrants in space of arousal and valence, and to groups of Ekman's basic emotions: **Happy**: positive arousal and positive valence, includes basic emotion happiness. **Calm**: negative arousal and positive valence, there are no basic emotions in this quadrant. Instead we include neutral. **Sad**: negative arousal and negative valence, includes basic emotions sad and bored. **Upset**: positive arousal and negative valence, includes basic emotions disgust, anger and fear.

Therefore, we can jointly use databases that are labeled in space of arousal and valence (Recola, Semaine), as well as those labeled discretely (EmoDB, Ravdess).

1.2 Learning from Features or Raw Audio
Traditionally in machine learning we first extract features from audio. This can be done using specialized software, such as OpenSMILE [3], or libraries, such as LibROSA [6].

With deep learning, it is possible to learn from raw audio sig-

nal. This recent approach is interesting, as in the raw audio signal there is encoded certain information that is missing in extracted features. Deep learning has two problems: (1) larger databases are needed for training, and (2) training is very computationally expensive, both regarding computational power and large amounts of memory needed.

2. DATABASES

There are many public audio databases available for use in affective computing. Most of them are targeted towards speech recognition or a subset of emotions, specific for a given problem (such as detecting frustration in call centers). We describe the few of them that we have used. [1]

We chose those based on the way they were labeled, language and audio format used. Regarding labels we preferred labels in space of arousal and valence or basic emotions in order to be able to do both regression in some emotional space or classification of emotions. We decided to only use European languages, since it has been shown that model trained on language from similar cultural background to target population gives slightly better results [2]. Audio simply needs to be of high enough quality. Human speech ranges up to 5kHz so we need at least 10kHz sampling rate. To be on the safe side and not lose any non verbal information we decided to only use audio recorded at 16kHz or higher.

2.1 EmoDB

EmoDB [1] (Berlin Database of Emotional Speech) is an older database. It contains 535 utterances spoken by 10 different actors. Each actor expressed each of the Ekman's 6 basic emotions (and a neutral version) at least once for each of the ten different texts. Each file is labeled. Texts themselves are emotionally neutral. Utterances are quite short, recordings are between a couple of seconds long up to half a minute.

Problematic aspects of this database are:

- Utterances are very short. Often when classifying audio, recordings are cut into segments from 1 second up. If we do that with EmoDB, there are simply not enough instances to use deep learning techniques, in same cases there are even not enough for traditional ML.
- Expressed emotions are extreme to the point of overacting. This means that classifiers trained on this set may produce weak generalization, as most speech is closer to neutral as considered in this database.

2.2 Semaine

The SEMAINE [7] database is a multi-modal database that includes audio, video and transcripts of English texts. The database is labeled on a continuous scale along many dimensions. Not all sessions (couple of minutes long recordings) are labeled in all dimensions. Most are labeled along the arousal and valence dimensions, as well as intensity and power. Fewer are labeled along basic emotions (e.g. only 2 session labeled for fear). Each available dimension is labeled by at least 2 annotators. Differences among different annotators are quite noticeable which is to be expected in such a setting.

Problematic aspects of this database are:

- Very unbalanced due to the chosen labeling methodology. Counting each label sample, there are almost 4x as many examples of low arousal and high valence than examples of high arousal and low valence.
- For some dimensions label values span a very small interval, which may cause problems with regression along those dimensions.
- Differences between annotators are often quite big. Some files have inter annotator correlations below 0.2. While this is not unexpected - emotion expression and perception are inexact - it is problematic for training and testing.
- Expressed emotions are very mild and often noticeably acted. There are examples in which we can hear the actor, supposedly gloomy and depressive, express amusement by laughing. While extreme emotions are problematic so are very mild emotions - ML algorithms often overfit to find other characteristics in the data.

2.3 Recola

Recola Database is a French multimodal datset of emotional speech. It includes audio, video, biosignals, labels (annotations) and metadata. It is similar to Semaine in that it is also labeled continuously. It is only labeled along arousal and valence dimensions, but labels are of higher quality. Each recording is labeled by 6 different annotators, 3 male and 3 female.

Problematic aspects of this database are:

- Each file is exactly 5 minutes long, but some of the labels are missing a few samples. We have cut the audio files to match the label lengths.
- There are only 23 recordings. Since each is 5 minutes long it is still quite large.
- It is quite unbalanced. Counted by each label sample, there are more than 8x as many examples of high arousal and high valence than examples of low arousal and low valence.

2.4 Ravdess

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains video and audio of 24 actors. Emotions expressed include the 6 Ekman's basic emotions, neutral and calm. Each utterance corresponds to one emotion. There are 1440 files in the speech section and 1012 in the song section.

Problematic aspects of the speech section of this database are:

- All utterances contain one of the two texts: (1) "Kids are talking by the door" or (2) "Dogs are sitting by the door". This may represent a problem as all files are in a way very similar. On the other hand, this is good as it helps prevent overfit as the algorithm can't learn to differentiate utterances based on text contained.

[1] Some of reportedly high quality databases such as the Humaine database and Vera am Mittag (eng. *Vera at noon*, a database of German emotional speech taken from reality TV and talkshows) are not available anymore.

- Utterances are very short, similar to EmoDB.

3. FEATURES

Among features used for classification of audio of human speech are (1) simple features such as loudness, signal energy and pitch, (2) Mel spectrum: similar to frequency spectrum, transformed to Mel scale which corresponds roughly to human perception of pitch, (3) Mel frequency cepstrum coefficients: inverse Fourier transform of log-scaled Mel spectrum, (4) Jitter, shimmer: frequency noise instability and amplitude instability, (5) Formants: most present harmonic frequencies, (6) Spectral features: describe the shape of the frequency spectrum and (7) Chroma features: describe tonal properties, such as melody.

3.1 Tools

Some of the commonly used tools for audio feature extraction are OpenSMILE and LibROSA.

OpenSMILE is a standalone program with a very step learning curve. Writing custom configuration files which is needed for extracting custom features as opposed to using one of the predefined features sets is quite complicated. Most users use predefined configurations, which can also be found online.

LibROSA library is an easy to use alternative that works with Python and offers similar functionality. It also offers some utility functions for reading and storing audio files, filters etc.

3.2 Feature Selections and Analysis

Feature selection is an important step in the ML pipeline as having fewer features is beneficial for reducing training time as well as reducing the possibility of overfitting.

We have performed feature selection using each of described databases, using features calculated by OpenSMILE (using a slightly modified ComParE13_LLD configuration) and separately using features calculated using the LibROSA library.

1. Remove features with variance below 0.2, as they hold little information.

2. Sort by correlation with labels and remove those with absolute correlation below 0.1 as they mostly contribute noise or bias towards groups with certain vocal qualities.

3. Greedy feature selection: take the feature with the highest correlation, add it to the feature set and test on a surrogate model (logistic regression or random forest classification). We use surrogates to reduce the computation time. Keep feature if it improves performance of the surrogate.

Using this method the number of features was reduced from 132 to 60 (feature set ComParE_lld extracted using OpenSMILE), and from 167 to 110 (custom feature set extracted using LibROSA). We achieved the same performance on the models while reducing the training time compared to no feature selection.

4. EXPERIMENTS

We have tested regression in the space of arousal and valence and classification of basic emotions in order to compare two very different approaches and decide which is preferable for our use-case.

4.1 Regression in Arousal and Valence Space

While deep learning on raw audio signal is slower and more computationally expensive, it may produce better results as raw signal contains more information. We have replicated the experiment done by Trigeorgis et al. [9]. Due to hardware constraints we had to introduce certain modifications: (1) we had to use 3 second segments instead of 6 second segments and (2) we used a mono-directional LSTM layers instead of bi-directional as in the paper. Network topology is otherwise same.

Training and testing was done on the RECOLA database. Data was split into train and test sets by actors - 80% of actors in the train set and 20% in the test set. Our results were very similar to those reported in the paper. Measurements shown in Table 1 are Concordance correlation coefficients (CCC)[2] between predictions and ground truth, obtained as averaged labels. Predictions are scaled to have the same standard deviation as the ground truth and time-shifted in order to remove any delays that a human annotator may produce. Thus we can confirm that deep learning from raw audio data is feasible.

	Arousal CCC	Valence CCC
Raw audio	0.641	0.250
Features	0.574	0.187
Trigeorgis et al. [9]	0.684	0.249

Table 1: Valence and arousal regression results

The network is made of two distinct functional units. First are the convolutional layers that learn to perform feature extraction. It has been shown [9] that certain neurons are highly correlated to some of the known good features. The second part is made of two LSTM layers. These learn to regress arousal and valence from extracted features.

The same experiment was repeated using only the second part of the neural network, trained using extracted features (feature set ComParE_lld). Results were somewhat worse, which indicates that the convolutional part of the full neural network learns to extract a better set of features than we get using simple feature selection (as described above). Unfortunately predictions in the space of arousal and valence are hard to interpret and there is no direct way to convert them to basic emotions.

4.2 Classification of Emotions from Features

We have used EmoDB for initial experiments. All reported results are averaged over leave-one-person-out cross validation. Simple fully connected feedforward neural networks tend to overfit. This can be reduced with hyperparameter adjustment (learning rate and algorithm, mini-batch size, early stopping etc).

[2]Concordance correlation coefficient is a measure of agreement, often used to measure inter-rater reliability.

This was tested on the database split into 3 sets, train, test and evaluation. Using 3 sets show that overfit is still there, but the difference in performance was small between train set and test set probably due early stopping based on test set loss. Performance on evaluation set is still much lower.

We used all features from the ComParE_lld feature set. Input layer therefore has 130 units, fist hidden layer 70, second hidden layer 30 and output layer 7 (6 Ekman's basic emotions + neutral). For the experiment, MSE was used as loss function, and Adam as optimizer. Without using regularization, we achieve very poor performance. As the model starts to overfit we stop training it, which is before it achieves good performance. Without regularization accuracy is therefore very low on all sets. Even with strong regularization, using both added Gaussian noise ($std = 1/2std(features)$) to input layer and dropout ($p = 0.5$), large differences on train set and evaluation set can be seen.

We compare out results to state of the art as achieved by Yenigalla et al. [10] in 2018 and Gjoreski et al. [4] from 2014. Yenigalla et al. achieved high performance using convolutional neural networks, trained using extracted features and phonemes. IEMOCAP dataset was used. Gjoreski used Auto-WEKA, a machine learning tool that automatically chooses best classical-ML algorithm. They trained and tested using EmoDB.

	Accuracy (test)	Accuracy (eval)
No regularization	0.54	0.48
Noise, dropout	0.82	0.65
Gjoreski et al.	/	0.77
Yenigalla et al.	/	0.73

Table 2: Classification results for test set and evaluation set, compared to state of the art [10].

We have also performed some preliminary experiments using Optimal Brain Damage (OBD) algorithm[5] to prune the network. Results are not yet conclusive but seem promising. We did not achieve better performance, but did achieve same performance while pruning up to 60% of all units.

5. CONCLUSION

We have experimented with regression in space of arousal and valence. Results confirm that a combined convolutional and recursive neural network can effectively learn on raw audio signal. Since the authors who propose this approach state that the convolutional part of the network learns to perform feature extraction we tested only the recursive part of the neural network, trained on pre-extracted features. Results were somewhat worse, which can be interpreted as the convolutional part of the network learns to extract better features. Aditional experiments, such as classification using a similar neural network are needed in the future.

We have also tried using a fully connected artificial neural network (FNN) to classify emotional speech. FNN is extremely prone to overfit. Even using very aggressive regularization techniques show some overfit. It seems that either (1) FNNs need a larger amount of labeled training data or (2) are not well suited for this problem. Related future work is performing experiments using OBD to prevent overfit.

A new categorization of emotions was proposed with the aim of using multiple databases jointly. Preliminary experiments show that we can use it for machine learning on multiple databases. Whether models trained in such way will perform better is yet to be seen.

In conclusion, emotion recognition using audio signal is a complex and difficult task. Some of our experiments come close to state of the art, but still not very good. We believe we can improve our work further in the future.

6. REFERENCES

[1] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[2] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[3] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[4] Martin Gjoreski, Hristijan Gjoreski, and Andrea Kulakov. Machine learning approach for emotion recognition in speech. *Informatica*, 38(4), 2014.

[5] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.

[6] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[7] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 3(1):5–17, January 2012.

[8] Venet Osmani, Agnes Gruenerbl, Gernot Bahle, Christian Haring, Paul Lukowicz, and Oscar Mayora. Smartphones in mental health: detecting depressive and manic episodes. *arXiv preprint arXiv:1510.01665*, 2015.

[9] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5200–5204. IEEE, 2016.

[10] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech emotion recognition using spectrogram & phoneme embedding. *Proc. Interspeech 2018*, pages 3688–3692, 2018.