Convoluted Feelings
Convolutional and recurrent nets for detecting emotion from audio data

Namrata Anand
namrataa@stanford.edu

Prateek Verma
prateekv@stanford.edu

Abstract

We present a Convolutional Neural Network (CNN) model to extract features from audio data. We apply this model to the task of classifying emotion from speech data. Classification on features extracted from our single-layer CNN model (1024 filters) outperforms classification on typically extracted acoustic features. We achieve an accuracy of 50% for 7-class classification using CNN-extracted features on 500ms audio patches. In addition, we predict emotion with 70% accuracy using LSTMs (Long Short Term Memory) to get context over time.

1. Introduction

With the near ubiquity of personal computing systems comes the demand for more naturalistic and personalized devices—devices which can listen to, process, and appropriately respond to human commands. In this paper, we try to train a deep learner to detect emotion from audio data, the goal being to build a system that can reliably detect human emotion either in an online or batch setting. Our title hints at the trickiness of the problem and at our determined approach.

Approach. Convolutional neural networks (CNNs) have the capacity to learn higher order data features, and as a computer vision tool, CNNs have proven to be extremely successful models for interpreting image data. We apply shallow and deep CNNs to audio spectrogram data, an image representation of audio data along frequency and time axes. We also experiment with fine-tuning a CNN to act as a simple facial expresion detector on images.

In order for a system to process audio data in an online fashion, it must be flexible enough to accept variable length input and to learn meaningful long-range temporal relations in the data. Recurrent Neural Networks (RNNs/ LSTMs) perform well on tasks that require integration of state information over time. Therefore, we also implement an LSTM on our dataset, in order to assess the importance of temporal context for emotion classification. Our work with recurrent models is just beginning; we hope to improve emotion classification through supervised training of a deep CNN/LSTM on audio spectrograms and image data as well (Figure 1). The CNN will detect relevant, complex features from short segments of speech data and the LSTM will integrate long-range feature patterns and classify emotions associated with audio in a sequential manner. This approach is inspired by the time fusion models in [6]. and the LRCN (Long-term recurrent convolutional nets) models implemented by Donahue et al. for image captioning [1].

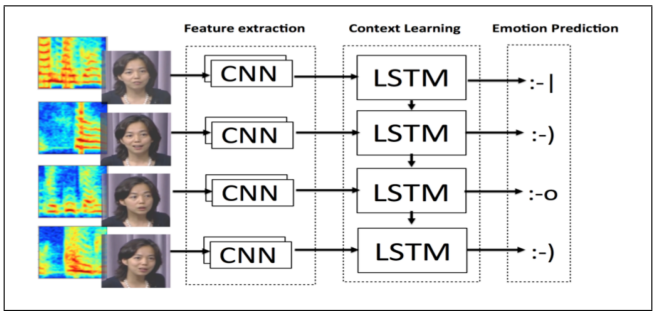

Figure 1. The full and final approach: feature extraction with CNNs on audio and image data; temporal context encoding with LSTMs (recurrent network). In this paper, we demonstrate that CNNs and LSTMs individually perform well for feature extraction and emotion classification.

Related work. CNNs have been used in the past specifically for the task of emotion recognition. Huang et al. attempted to learn salient feature maps for the task of emotion recogniton using an auto-encoder followed by a CNN and achieved good performance [4]. However, no temporal model was used; a linear classifier was trained across features from all time frames in a given audio file to predict the associated emotion. We believe this approach is not flexible as it does not allow for variable length input audio.

Taking temporal context into account, Graves et al. implemented a bidirectional LSTM on hand-built audio features and demonstrated the importance of context learning for the task of emotion recognition [2]. We believe we can improve on these results by extracting features with a CNN—in fact, for this particular classification task, our top

CNN model strongly outperforms classification on the same baseline acoustic features used in [2].

Since we are working with an amalgamation of many data sources, it is not clear what result is the "state-of-the-art" with respect to the data, or if even reporting such a thing makes sense given the small size of the dataset. Therefore, we cannot make direct comparisons with other results, although these are useful benchmarks. Li et al. were able to achieve 53% test set accuracy on the eNTERFACE'05 dataset using a hybrid 6-layer Deep Neural Network–Hidden Markov Model on hand-built features similar to those described in the next section [7]; our CNN model achieves 55% accuracy on our augmented version of the eNTERFACE'05 dataset– *without taking any temporal context into account*.

One successful and somewhat similar approach by Kahou et al. (2013) on an extensive dataset[1] involved using CNNs on face image data and DBN (Deep Belief Nets) on audio features. They achieved a highest test set accuracy of 47.7% for a 7-class classification problem on a larger dataset (lim 2000 sequences ranging from 300-5400ms in length) [5]. However, in our approach, we try to replace hand-built features entirely and use CNNs to extract features from both spectrogram as well as image data.

2. Data

Datasets. We use data from three emotional speech databases to train and validate our models. These are described in the table below.

- The Berlin Emotional Speech dataset consists of 537 utterances in German of 10 unique statements by 10 different actors portraying 7 emotions—happy, angry, sad, disgusted, fearful, bored, and neutral. The actors are native speakers. The files are on average 2-3 seconds long.
- The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset consists of 480 British English utterances recorded by 4 male actors portraying 7 emotions—happy, angry, sad, disgusted, fearful, surprised, and neutral. The files are on average 4-5 seconds long.
- The eNTERFACE'05 emotion dataset consists of audio data from 42 subjects of 14 different nationalities. The speech data was pulled from video files of the subjects speaking in English. The subjects portray 6 emotions—happy, sad, angry, disgusted, fearful, and surprised. The files are on average 3-4 seconds long.

[1]The dataset used here was offered as part of the Emotion Recognition in the Wild competition. The competition had closed, so we couldn't get access to the dataset.

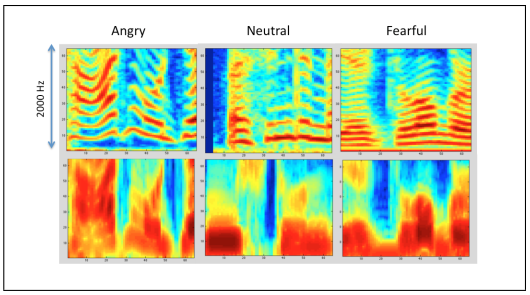

Figure 2. Example spectrogram data. 512ms chunks from audio data files corresponding to three distinct emotions. Two channel representation– top channel represents pitch modulation and bottom channel represents formant changes.

Each dataset is balanced across emotion classes, but the classes are not balanced across the dataset as a whole.

Data preprocessing. We resampled the data to 16000 Hz and generated spectrogram representations of the audio in MATLAB by taking the magnitude of the Fast Fourier Tranform (FFT) with 2048 points. Each spectrogram has two channels, the first with frame size of 40 ms, the second with frame size 5 ms, and both with 2 ms hop. Both channels encode similar information, but the first channel largely represents pitch modulations (higher frequency resolution, lower time resolution) while the second represents changes in timbre (lower frequency resolution, higher time resolution).

We then log-transformed and subsampled the spectrograms to frequency range 1-2000 Hz and split them into non-overlapping chunks each representing 512 ms of speech, assigning each chunk the emotion label attributed to the entire audio file. This gave us a total of 22,521 input data spectrgrams. This approach leads to some degree of data loss, and we hope to refine this process for later models. We resized the spectrogram images to a final resolution of 64x64x2. Example chunks for audio file can be seen in Figure 4.

Feature	Num.	Description
RMS energy	1	Volume across time
MFCC coeff.	13	Timbre characterized by 13 coeff.
Zero crossing	1	Noise in sample
Spectral energy	4	Energy in bands—formants of signal
Roll-off	5	Tonal quality measurment
Centroid	1	Description of timbre
Pitch	1	Fundamental freq. of sound
Total feat.	26	

Table 1. Audio-extracted features

We also extracted 26 hand-built features from the audio files for each 512 ms chunk. These are features that have been used in past studies to identify emotion from speech data [2]. The features are described briefly in Table 1 and are described in detail in the paper's appendix.

We identified and removed mislabeled examples in our input data vector (about 0.5% of values) and replaced anomalous data points with the median of 1000 randomly selected points from the dataset. We concatenated all three datasets, shuffling the order of data, performed mean subtraction and assigned 60% to the training set, 20% to the validation set, and 20% to the test set (For experiments without a validation set, we split the training and test set at 70% and 30%, respectively). Due to relatively low speaker diversity in our dataset, our training and test sets are not speaker independent, somewhat limiting the generalization ability of our models.

3. Results

3.1. Audio feature extraction

Results are outlined in Table 2 and described below.

Model	Train acc.	Val acc.	Test acc.
Hand-built feat. SVM	0.264	0.25	**0.237**
Hand-built feat.softmax	0.245	0.227	**0.212**
FC 2 layer NN	0.327	0.274	**0.273**
FC 3 layer NN	0.301	0.2614	**0.255**
Best sq filter CNN (S2)	0.820	–	**0.364**
Best rect filter CNN (R1)	0.720	–	**0.397**

Table 2. Audio feature extraction results for linear models, fully connected neural nets, and CNNs

Baseline models We ran SVM and softmax to classify the input data (512 ms) based on hand-built features described in Table 1. We did coarse hyperparameter tuning to find optimal learning rates and regularization. The best SVM and softmax models achieved 23.7% and 21.2% classification accuracy on the test set, respectively. We then implemented a two-layer fully connected neural network with a softmax classifier on top. We did coarse hyperparameter tuning over learning rate, regularization, number of units in hidden layer, and number of epochs. The net run with tuned parameters showed higher performance than linear classifiers on the hand-built features (Table 1). The best model achieved 27.3% classification accuracy on the test set. Learned weights for the best model looked like spectrograms but were difficult to interpret. We increased the depth of the fully connected neural network by adding an additional hidden layer. The net showed higher performance than linear classifiers on the hand-built features but lower performance compared to the single hidden layer network. (Table 1). The best model achieved 25.5% classification accuracy on the test set.

CNNs. In order for CNNs to learn meaningful representations from spectrogram data, we must not convolve over the frequency axis– that is, our convolution filters must be rectangular, with height exactly equal to the height of the image. In practice, however, others have implemented square filters and have reported similar results between 1-D and 2-D convolution.

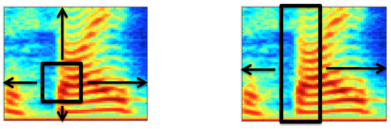

Figure 3. Representation of 1-D vs. 2-D convolution across a spectrogram image.

We experimented with a number of 1- and 2-layer CNNs with both square and rectangular filters in an effort to get the highest classification accuracy for 512 ms patches of audio. Generally, we saw a drop in performance with deeper CNNs.

Optimal model architectures and hyperparameters were selected by randomized grid search and nets were tuned for 10 epochs. The top rectangular and square models were then trained for 400 epochs. Results are reported in Tables 1 and 4. The strongest single model that we tested on the original dataset was a one layer convolutional network with 1024 64x5 filters (R1), achieving a test accuracy of 39.7%. Overall, we saw that models with rectangular filters seemed to outperform those with square filters slightly. Widening filters, adding dropout, and using leaky ReLUs did not improve model performance. We extracted class scores from R1 and attempted to classify emotions on these with simple linear models (Table 3). This also did not change performance, but these serve as a nice baseline for higher order models that might build off the CNN-extracted features.

Model	Test acc
Nave Bayes	0.386
Softmax	0.396
LDA	0.388
QDA	0.391
GMM (7 mixtures)	0.392

Table 3. Baseline models trained on class scores for top CNN model (R1)

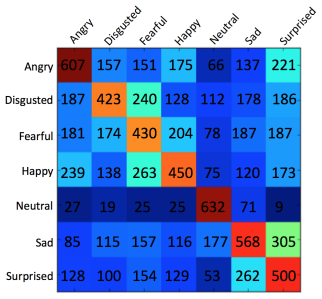

Figure 4. Confusion matrix for highest performing CNN model (R1). Emotions represented in following order: angry, disgusted, fearful, happy, neutral, sad, surprised.

Model	Test acc	Architecture	Filter size	Hidden units
Square Filters				
S1- Conv16	0.351	1 (conv-relu-pool-[affine]x2)	5x5	64,7
S2- Conv32	**0.364**	1 (conv-relu-pool-[affine]x2)	5x5	64,7
S3- Conv16/32	0.355	2 ([conv-relu-pool]x2-[affine]x2)	5x5	64,7
S4- Conv256	0.368	1 (conv-relu-pool-[affine]x2)	5x5	128, 7
Rect. Filters				
R1- Conv1024	**0.397**	1 (conv-relu-pool-[affine]x2)	64x5	64,7
R2- Conv32	0.368	1 (conv-relu-pool-[affine]x2)	64x5	64,7
R3- Conv16/32	0.339	1 (conv-relu-pool-[affine]x2)	64x5	64,7
R4– Conv128	0.334	2 ([conv-relu-pool]x2-[affine]x2)	64x10	64,7
R5– R1+ wide filters	0.394	1 (conv-relu-pool-[affine]x2)	64x10	64, 7
R6- R1+Dropout, leaky ReLUs	0.380	1 (conv-relu-pool-[affine]x2)	64x5	64,7
R7- R1+ data aug. (eNTERFACE)	**0.550**	1 (conv-relu-pool-[affine]x2)	64x5	64,7
ER1- R1 ensemble over time	0.392			
ER2- Best Ensemble (R1+R3)	**0.408**			

Table 4. Performance of top CNN models for emotion classification on audio data: 1- and 2-layer convnets with varying convolution filter sizes and hidden units in fully connected affine layer. Nets were trained for 400 epochs.

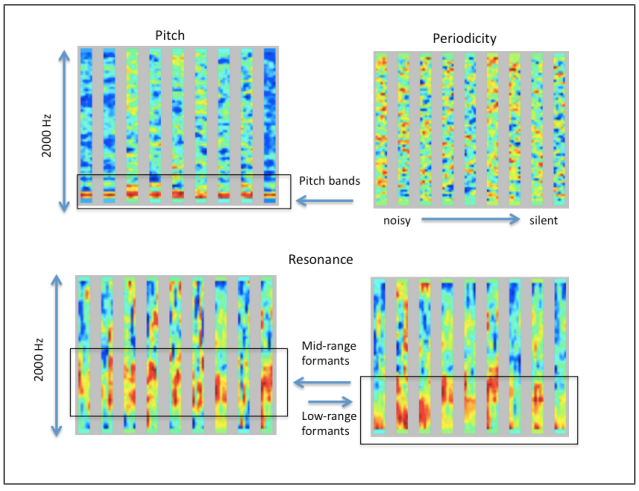

Figure 5. Weight visualization of top CNN (R1) features. CNN discovers typically used hand-built features in spectrogram processing, including pitch, resonance, and periodicity information.

Weight visualization

We visualized the first layer convolution weights from our top rectangular filter model (R1). At a high level, we noticed that the filters seemed to be picking up on features that are typically extracted from audio data (e.g. Table 1). We used hand-built features as metrics to select the top filters corresponding to different pitches, high and mid-range formant sensitivity, periodicity (an indicator of silence, fricatives, breathiness, noise, etc.), and silence. Visualization of some particularly indicative weights can be seen displayed in Figure 5.

Data augmentation

Our models display high variance, which is likely a function of the relatively small training set size. We attempted to ameliorate the overfitting with regularization, but the generalization capability of even our best model is extremely limited.

With the hopes of reducing generalization error, we augmented our dataset (9X) with pitch shifts and time scaling. We saw immediate increase in the performance of our top model, with highest validation accuracy of 55% (R7). The model also seems to generalize better and displays reduced overfitting.

3.2. Emotion recognition from face images

Model	Train acc.	Test Acc
Softmax	0.4	0.305
FC 2 layer NN	0.5	0.321
Fine-tuned Caffenet	**0.92**	**0.886**

Table 5. Fine-tuned deep CNN achieves high classification accuracy on face images

Many promising deep learning results on facial recognition including [1] indicate that visual cues might supplement emotion detection from audio. We curated a training set of facial expression images from the eNTERFACE'05 dataset by sampling 5 random image frames from videos of the subjects. The training data is composed of centered, front-facing faces against a neutral background. We used these images to finetune a deep CNN trained on the ImageNet dataset ("CaffeNet"). The fine-tuned CNN greatly outperforms baseline models, but due to the small size and specific nature of images in the dataset, does not generalize very well. Results are reported in Table 5.

3.3. Recurrent networks– context learning

LSTMs. We hope to classify audio data in real time using LSTMs, which have been shown to classify emotion from speech data well [2]. We first ran experiments on our raw spectrogram data, piping in non-overlapping audio frames of 512ms. Since we would eventually like to combine audio-visual cues, we only used data from eNTERFACE here—1250 audio files, each around 3-5ms. The best LSTM model achieved **70%** test set accuracy and

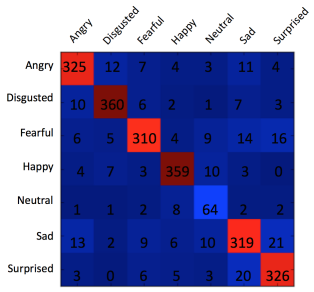

Figure 6. Confusion matrix for fine-tuned Caffenet model. Emotions represented in following order: angry, disgusted, fearful, happy, neutral, sad, surprised.

87.5% training accuracy after training for 50 epochs (Figure 7). We expect this performance to increase when running the LSTM on extracted features using our top CNN model. Since deep CNN/LSTM(RNN) architectures have performed well on sequential visual description tasks [1], we hope to integrate the best models (best CNN/ best LSTM) into a deep CNN/LSTM model for online emotion classification of audio data. For batch classification, it might also be interesting to try out bidirectional LSTMs here which have worked well for speech recognition tasks [3].

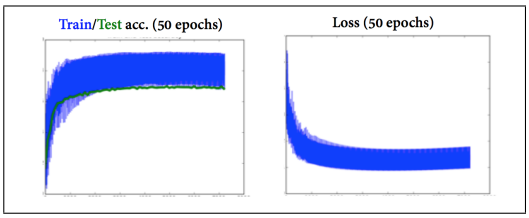

Figure 7. Learning and loss curves for LSTM model. Model shows less variance than CNN model.

4. Discussion

In this paper, we show that CNNs are a reliable method of feature extraction for audio data. Our highest performing models recover many of the typical hand-built features used to characterize spectrograms. We found that using rectangular filters for convolution (1D) led to slightly better performance over square filters (2D). We also found that the shallow CNNs (1/2 layers) with many filters performed better than deeper representations with fewer filters.

Our best model classifies the emotion of audio patches of 512ms into one of seven categories with 55% accuracy. Still, CNNs on brief patches of audio data do not suffice, as they cannot pick up on larger context of speech over time; this is reflected in the high variance of the models. LSTMs run on non-overlapping 512ms patches extracted from full audio files in our dataset (around a few seconds in length) significantly outperform CNNs in terms of classification ability and seem to generalize well.

Some issues with our approach that must be stated— We are working with a small, albeit diverse, dataset of short snippets of (sometimes poorly acted) emotional speech. As a result, we cannot definitively speak to the generalization ability of our models or the consistency of our results across different kinds of audio data (noisy, multiple speakers, etc.). Our emotion classes were not entirely balanced, but we were ill-disposed to throwing out training data.

Our next step is to integrate feature extraction and context learning into a single system which can accurately predict emotion from audio data of variable length in an online manner. First, we will study LSTM performance on CNN-extracted features. If high-performing, this will be a "deploy-ready" system, with a trained CNN and a trained LSTM as disparate parts. After procuring more data, we will attempt to train a model like the one pictured in Figure 1—where feature extraction and context learning from both audio and visual cues are united into a single model.

References

[1] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[2] Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. Online emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1-2):7–19, 2010.

[3] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005*, pages 799–804. Springer, 2005.

[4] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the ACM International Conference on Multimedia*, pages 801–804. ACM, 2014.

[5] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Aaron Courville, Pascal Vincent, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *arXiv preprint arXiv:1503.01800*, 2015.

[6] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014.

[7] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 312–317. IEEE, 2013.