INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms

Aharon Satt[1], *Shai Rozenberg*[1, 2], *Ron Hoory* [1]

[1]IBM Research-Haifa, Israel
[2]Technion-Israel Institute of Technology, Israel
aharonsa@il.ibm.com, shayr@il.ibm.com, hoory@il.ibm.com

Abstract

We present a new implementation of emotion recognition from the para-lingual information in the speech, based on a deep neural network, applied directly to spectrograms. This new method achieves higher recognition accuracy compared to previously published results, while also limiting the latency. It processes the speech input in smaller segments – up to 3 seconds, and splits a longer input into non-overlapping parts to reduce the prediction latency.

The deep network comprises common neural network tools – convolutional and recurrent networks – which are shown to effectively learn the information that represents emotions directly from spectrograms. Convolution-only lower-complexity deep network achieves a prediction accuracy of 66% over four emotions (tested on IEMOCAP – a common evaluation corpus), while a combined convolution-LSTM higher-complexity model achieves 68%.

The use of spectrograms in the role of speech-representing features enables effective handling of background non-speech signals such as music (excl. singing) and crowd noise, even at noise levels comparable with the speech signal levels. Using harmonic modeling to remove non-speech components from the spectrogram, we demonstrate significant improvement of the emotion recognition accuracy in the presence of unknown background non-speech signals.

Index Terms: Speech Emotion Recognition, Para-lingual, Deep Neural Network, Spectrogram.

1. Introduction

Emotion recognition solutions are becoming one of the latest trends in the global IT market [1]. These advanced solutions capture and analyze the human emotions from multiple sources, where the human voice serves as one of the major ones. Emotion recognition capabilities are required to support natural and efficient human-computer interaction, generate marketing insights, help discovering entertainment content across large repositories and playlists, enable effective e-learning through e-tutors, and many more [1-3].

Emotion recognition from the para-lingual information in the speech has gone through a significant improvement over the recent years, with the introduction of deep neural networks. Yet many challenges still hold, such as the need to keep improving the recognition accuracy, reducing its latency and enhancing its robustness to background sounds – from music to everyday noises. Human-computer interaction, for example, requires low-latency emotion recognition, possibly with everyday background noise, to support a dialog on the fly. Media and entertainment content discovery requires handling of speech with music and other media-related background sounds.

In this work we present our approach for dealing with these challenges. We follow the recent success of applying deep learning methods directly to spectrograms, across different areas of speech processing, to design our solution.

This paper is organized as follows: the relevant related work is surveyed in section 2; the proposed solution is described in section 3, and its evaluation results are presented at section 4; section 5 describes as extension to handle effectively loud background signals; finally, section 6 presents the conclusions.

2. Related work and evaluation dataset

Emotion recognition from the para-lingual component of the speech has been an active research area for decades [3-8]. Traditional methods were based on short-time frame-level feature extraction, followed by utterance-level information extraction, and classification or regression as required [3-8]. In the recent years, deep learning methodologies and tools have been introduced to this area, used for feature extraction, classification/regression, or both [9-14].

Evaluation and comparison of different solutions is challenging, as no sufficiently comprehensive labeled public corpus of emotional speech is yet available [3-8]. One of the best available corpora is IEMOCAP [15]. It contains a relatively large amount of data, labeled down to the single sentences. State of the art results have been published using this corpus [9-10]. For this reason, we chose the IEMOCAP dataset for the evaluation of our proposed system.

Deep learning has contributed to breakthroughs across multiple areas of multimedia, including speech processing [16-17]. Researchers have shown that replacing hand-crafted low-level (frame-level) features with statistical learning by the different layers of the deep network can significantly enhance the accuracy of detectors and regression solutions. In speech recognition, the shift back from MFCC representation to Mel-scale spectrograms [16, 18] led to a reduced error. Direct use of Mel-scale spectrograms for speaker recognition was proved successful as well [19]. In [20-21] a recently published state of the art robust speech recognition system is described based on linearly-spaced spectrograms. In the present work we follow this path.

For speech recognition, Mel-scale spectrograms that eliminate part of the pitch information, can give rise to a superior solution. Conversely, emotions are strongly manifested in the pitch information; hence we turned to design our system using linearly-spaced spectrograms, which represent the fine harmonics structure of the speech. A recent publication [22] demonstrates the use of such spectrograms to remove

http://dx.doi.org/10.21437/Interspeech.2017-200

background music and deliver the clean speech. Our aim is to design a high-accuracy emotion recognition system from speech, while limiting its latency and enhancing its robustness to background sound signals, in particular loud signals.

3. The proposed system

3.1. Speech input processing and spectrogram calculation

As described above, we calculate spectrograms from the speech signal and apply deep learning directly to the spectrograms. The speech signal in the IEMOCAP corpus [15] is sampled at 16KHz and organized as single sentences with durations from less than a second to about 20 seconds. Each sentence is labeled with one emotion.

As the first step, we split each sentence longer than T=3 seconds to shorter sub-sentences of approximately equal lengths, not longer than T=3 seconds. Each sub-sentence is assigned the emotion labeling of the corresponding whole sentence. These shorter sentences are used throughout the proposed system, where only during the testing phase we evaluate the prediction for the whole sentences by averaging the posterior probabilities of the respective sub-sentences. While we lose some accuracy in this process, our aim is to propose a system that limits the prediction latency.

Next, we calculate a spectrogram for each (shorter) sentence. A sequence of overlapping Hamming windows is applied to the speech signal, with frame size (window shift) of 10msec, and window size of either 20msec or 40msec. For each frame we calculate a DFT of length 800 (for 20Hz grid resolution) or 1600 (for 10Hz grid resolution). We use the frequency range of 0-4KHz, ignoring the rest. Following aggregation of the short-time spectra, we obtain a matrix of size NxM, where N<=300 according to the speech sentence length, and M=200 or 400 according to the selected frequency grid resolution.

Next, we implement a normalization step: the DFT data is converted to log-power-spectrum, expressed in dB; it is then limited from below by the constant E_{noise} that was determined empirically to represent a universal noise level; the resulting log-spectrum was lifted to be non-negative by subtracting the constant E_{noise}, and then normalized to bring its non-zero data points to a unity variance.

The last step in calculating the log-spectrogram is zero-padding to get 300 time points.

3.2. The deep neural network

We chose to evaluate two types of neural networks: convolutional networks and recurrent networks, where the latter refers to an LSTM – Long Short Term Memory networks. Figure 1 depicts an example of the deep network.

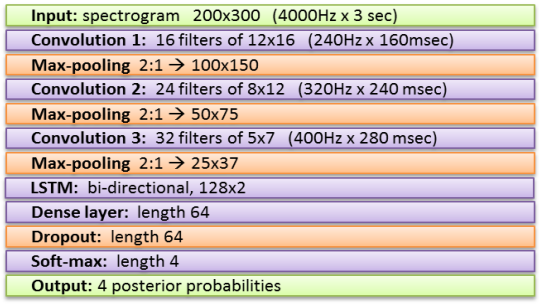

Figure 1: *Example of the deep network topology*

We hypothesized that the convolutional networks, capable of learning spatial patterns, will learn effectively spatial spectrogram patterns that represent the emotional information. We provide visual insights for this in the next section.

We also hypothesized that adding an LSTM layer will help learning the temporal behavior across the sentence being represented by the spectrogram. This hypothesis is supported by the improved accuracy as presented in the next section.

4. Experimental setup and evaluation

The IEMOCAP corpus comprises five sessions; each session includes labeled emotional speech sentences from recordings of dialogs between two persons. There is no speaker overlapping between different sessions. We used this setup for running a five-fold cross validation. In each fold, the data from four sessions is used for training the deep neural network, and the data from the remaining session is split – one speaker for validation and the other for the accuracy testing.

The IEMOCAP corpus contains scripted and improvised dialogs. As the script text exhibits strong correlation with the labeled emotions, it may give rise to lingual content learning, at least partially, which is an undesired side effect. Therefore we used the improvised data only.

We used two common evaluation criteria:

1. **Overall accuracy** – where each sentence across the dataset has an equal weight, AKA **weighted accuracy**;
2. **Class accuracy** – the accuracy is first evaluated for each emotion and then averaged, AKA **unweighted accuracy**.

For the sake of comparison to [10], the following four emotions were used: Anger, Happiness, Neutral and Sadness.

We tested dozens of combinations of topologies and parameters. We evaluated convolution-only topologies, ranging from two to eight layers, with different combinations of time windows sizes and frequency grid resolutions. We also evaluated topologies with one to six convolution layers and with one and two LSTM layers. The following table summarizes the best topologies, convolution-only and convolution with LSTM.

Table 1: *Summary of accuracy evaluation results based on five-fold cross validation*

	Network	Overall accuracy	Class accuracy
1.	5 convolution layers 10Hz grid resolution	66.1%	56.6%
2.	5 convolution layers 20 Hz grid resolution	63.6%	53.4%
3.	3 convolution layers and LSTM 10 Hz grid resolution	**68.8%**	**59.4%**
4.	3 convolution layers and LSTM 20 Hz grid resolution	65.8%	56.6%
5.	Two-step predictor (refer to explanation below) 10 Hz grid resolution	**67.3%**	**62.0%**

With respect to Table 1 above, we used the following parameters:

1. The window size was set to 40msec; a 20msec window yielded similar results, lower by 0-2% across the different topologies;

2. The bi-directional LSTM contained 128x2 nodes; using 64x2 nodes, the accuracy drops by 1-3%;
3. The frequency grid resolution was set to 10Hz; lower resolution (20Hz) yields lower accuracy by 1-3%;
4. The best topology for convolution-only network was found to include 5 layers (we tried 2-8 layers), whereas the best mixed-topology was found to include 3 convolution layers and a single LSTM layer (we tried 1-6 convolution layers and 1-2 LSTM layers);
5. The deep network was optimized to maximize the overall accuracy (this is discussed below).

The published state of the art accuracy using the IEMOCAP corpus, to our knowledge, is given in [10], based on the same evaluation setup as we used; it reports **63.9%** and **62.8%** for the overall accuracy and the class accuracy, respectively. It should be noted that [10] and other works present accuracy results based on the whole speech sentences; conversely, we split the sentences into shorter sub-sentences of T<=3 seconds, demonstrating the accuracy under limited latency constraint.

The IEMOCAP corpus is significantly unbalanced; to cope with the unbalanced data we tried the following techniques:

1. Training the network to maximize the class accuracy rather than the overall accuracy → the penalty on the overall accuracy makes it less useful;
2. Assigning different weights to the stochastic gradient, in inverse proportion to the class size → it improved both the overall and the class accuracies by 1-3%;
3. Applying statistical oversampling to get equal-sized training classes → increased the smallest class accuracy (happiness), but not the overall and class-accuracies.

We also tried a two-step prediction, based on:

1. A four-class predictor as in Table 1/row 3;
2. Three two-class predictors, which classify the majority class (neutral) against each of the other three classes; they use convolution layers as in Table 1/ row 1.

The two-step prediction process proceeds as follows: first, the test sample is run through the four-class predictor; if the higher probability is assigned to a non-majority class (non-neutral), then this class is selected; otherwise, the test sample is run through the three two-class predictors, and the predicted emotion is selected to maximize the posterior probability across the three predictors. The obtained accuracy using this two-step procedure is shown in Table 1/line 5, demonstrating higher class-accuracy. A heuristic explanation for the success of this method (special emphasis on the neutral class) could be due to the fact that significant parts of a typical non-neutral sentence tend to be neutral, whereas the emotional (non-neutral) nature is typically manifested in the smaller parts.

It is informative to examine what the deep network learns, by looking at the activations of the convolution layers. The Figures below show the activations of select filters at the first convolution layer, from a speech sample labeled as neutral.

Figure 2 – the left side – shows the original normalized log-spectrogram. The horizontal axis denotes the time, and the vertical – the frequency. Figure 2 – the right side – shows the activation of one of the filters. Reddish colors designate high activation, and blueish colors low activation. It is clearly seen that this filter tends to learn vertical and close-to vertical patterns of the fine harmonic structure in the log-spectrogram.

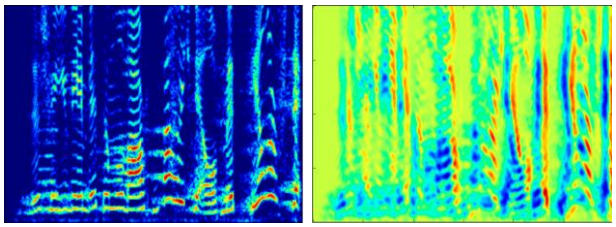

Figure 2: *Left: original log-spectrogram; right: activation*

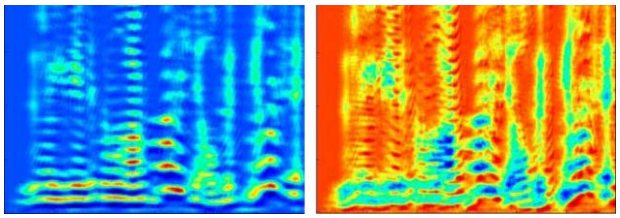

Figure 3: *Left and right – activations learn patterns*

Figure 3 – the left side – shows the activation of another filter, which clearly tends to learn horizontal and close-to horizontal patters of the harmonic structure.

Figure 3 – the right side – demonstrates a filter that tends to learn the less-relevant areas of the spectrogram, including silence and low-energy zones. This activation explains how the deep network is capable of separating the relevant parts of the spectrogram from the less-relevant areas.

To further enhance the recognition accuracy of the proposed solution, we tried to add a unidimensional attention mechanism to the LSTM layer. Our motivation, based on the success of bi-dimensional attention mechanisms in object recognition from images [25-26], was to find the time segments of the speech signal that are relevant to emotion recognition. Unfortunately, we have not gained any improvement of accuracy, concluding that in our case the convolution and LSTM layers seem to detect the relevant time segments effectively from the log-spectrogram, by themselves.

5. Emotion recognition in noise conditions

As discussed in section 1, several key use cases require the capability of emotion recognition from speech in the presence of background signals, such as music or crowd noise. The magnitude of these background signals can be close to or the same as the magnitude of the speech itself.

5.1. Background

A common way to filter out the unwanted signals (noise) and retain the speech information is based on de-noising auto-encoders [27, 22]. This method is useful if the noise characteristics are known, or it can be represented by a large enough set of samples.

In the current work, however, we chose to demonstrate a different method: removing the noise from the log-spectrogram itself, as a pre-processing step. Its significant advantage stems from the fact that no prior knowledge about the noise signal is needed, other than one constraint: it should not be a short-time harmonic signal, in the sense of exhibiting a single dominant short-time fundamental frequency within the pitch frequency range of humans.

In addition, we chose to demonstrate a method capable of handling high noise levels, as high as 0dB signal to noise ratio.

Emotion recognition from speech under lower noise levels was demonstrated at [7]: Gaussian noise with SNR = –20dB.

5.2. The noise filtering

We base our method on the short-time harmonic nature of the voiced speech. We made an assumption that the emotion information is contained mainly in the voiced parts of the speech. It is an approximation, but a useful assumption as shown below.

First, we use a common open-source pitch detector [23-24], to evaluate the pitch frequency of the speech, per frame. Then, for each voiced frame we generate a modified log-power-spectrum, which empirically approximates the real one:

$$S(f) = E(f) - 0.5\cdot(1 - \cos(2\pi f / F_0))\cdot D(f) \quad (1)$$

Where $S(f)$ is the modified short-time log-power-spectrum,

$E(f)$ is the short-time spectral envelope,

F_0 is the pitch frequency,

$D(f)$ is linear from 20dB @0Hz to 12dB @4KHz,

f is the frequency, $0 <= f <= 4KHz$.

Unvoiced frames are regarded as silence. Figure 4 depicts an example of the modified log-power-spectrum.

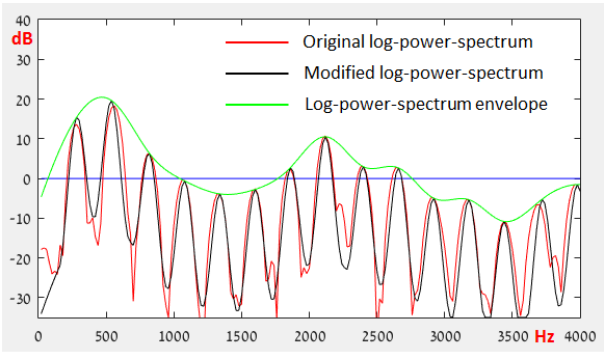

Figure 4: *Original and modified log-power-spectra*

In Figure 4 it is seen that the harmonic structure is approximately preserved, while any non-harmonic component would be largely eliminated. By replacing the normal log-power-spectra with their corresponding modified versions, we get the modified log-spectrogram.

5.3. Evaluation

For evaluating the noise immunity, we used seven different noise signals:

- **Three music signals:** passages from ambient, pop and rock compositions; all excluding singing;
- **Four crowd noises:** crowd speech, crowd anger sound, crowd applause and crowd laughter.

For testing we kept one fold (IEMOCAP session) aside, and trained two predictors based on the remaining four sessions:

1. The **normal predictor**, based on 3 convolution layers and LSTM as listed in Table 1/row 3;
2. A **modified predictor**, similar to the normal one but trained based on the modified spectrograms of the clean training data.

Next, we added (separately) the seven noise signals to each of the testing speech samples. The noise level was set to be at 0dB signal to noise ratio.

Finally we run the noisy samples through the two predictors:

- Noisy signal → noisy spectrogram → normal predictor,
- Noisy signal → modified spectrogram → modified predictor.

Calculating the modified spectrograms required extracting the pitch from the noisy signals, using the open-source [23-24].

Table 2 below summarizes the prediction accuracy.

Table 2: *Accuracy evaluation results for noisy speech*

	Test data	Predictor	Overall accuracy
1.	Clean spectrograms	Normal	68.8%
2.	Modified spectrograms from clean speech	Modified	64.5%
3.	Noisy spectrograms	Normal	40.1%
4.	Modified spectrograms from noisy speech	Modified	59.8%

As seen in Table 2, the spectrogram modification process takes its toll on the accuracy – comparing rows 1 and 2.

The value of the noise filtering becomes apparent when testing the noisy speech: comparing rows 3 and 4, the normal predictor's accuracy collapses, whereas the impact on the modified predictor's accuracy is relatively minor.

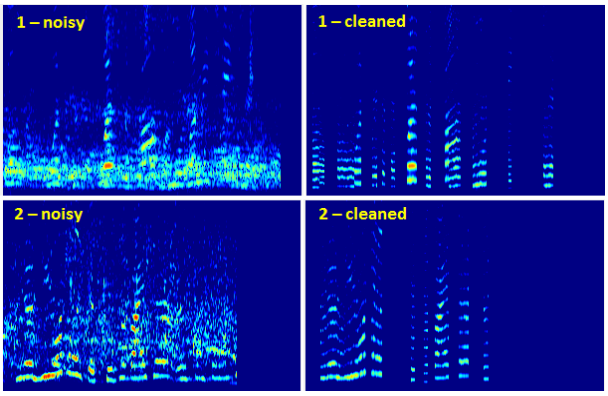

Figure 5: *Left: noisy spectrograms; right: cleaned*

6. Conclusions and discussion

We demonstrated an emotion recognition system from speech, under limited latency constraint (<=3 seconds), which achieves beyond the state of the art accuracy on the common benchmarking dataset IEMOACP, comparing with previous works without latency constraints: one of the tested network topologies achieves 67.3% and 62.0% vs. previous work – 63.9% and 62.8%, for overall- and class-accuracy, respectively. The system is based on an end-to-end deep neural network, applied directly to raw spectrograms without a feature extraction step. Using raw spectrograms enables us to easily combine a noise reduction solution based on harmonic filtering, which can handle high noise levels such as SNR=0dB – we demonstrated robustness to this level in the case of background non-speech noise sounds.

7. Acknowledgements

The authors wish to thank Mr. Izhak Golan from the Technion-Israel Institute of Technology, for his contribution for this research.

8. References

[1] Market research report: “Emotion Detection and Recognition Market by Technology, Software Tool, Service, Application Area, End User, and Region - Global Forecast to 2021”, Markets and Markets, November 2016.

[2] Market research report: “How To Measure Emotion In Customer Experience”, Forrester, November 2015.

[3] Vinola, C., and K. Vimaladevi. "A survey on human emotion recognition approaches, databases and applications." *ELCVIA Electronic Letters on Computer Vision and Image Analysis* 14.2 (2015): 24-44.

[4] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition* 44.3 (2011): 572-587.

[5] Chandrasekar, Purnima, Santosh Chapaneri, and Deepak Jayaswal. "Automatic speech emotion recognition: A survey." *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on*. IEEE, 2014.

[6] Schuller, Björn, et al. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge." *Speech Communication* 53.9 (2011): 1062-1087.

[7] Huang, Zhengwei, et al. "Speech emotion recognition using CNN." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.

[8] Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion recognition from speech: a review." *International journal of speech technology* 15.2 (2012): 99-117.

[9] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." *Interspeech*. 2014.

[10] Lee, Jinkyu, and Ivan Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition." *INTERSPEECH*. 2015.

[11] Huang, Zhengwei, et al. "Speech emotion recognition using CNN." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.

[12] Le, Duc, and Emily Mower Provost. "Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks." *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013.

[13] Rana, Rajib. "Emotion Classification from Noisy Speech-A Deep Learning Approach." *arXiv preprint arXiv:1603.05901* (2016).

[14] Chernykh, Vladimir, Grigoriy Sterling, and Pavel Prihodko. "Emotion Recognition From Speech With Recurrent Neural Networks." *arXiv preprint arXiv:1701.08071* (2017).

[15] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42.4 (2008): 335.

[16] Deng, Li. "A tutorial survey of architectures, algorithms, and applications for deep learning." *APSIPA Transactions on Signal and Information Processing* 3 (2014): e2.

[17] Deng, Li, and Dong Yu. "Deep learning: methods and applications." *Foundations and Trends® in Signal Processing* 7.3–4 (2014): 197-387.

[18] Deng, Li, et al. "Recent advances in deep learning for speech research at Microsoft." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[19] Variani, Ehsan, et al. "Deep neural networks for small footprint text-dependent speaker verification." *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[20] Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." *arXiv preprint arXiv:1412.5567* (2014).

[21] Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." *arXiv preprint arXiv:1512.02595* (2015).

[22] Simpson, Andrew JR, Gerard Roma, and Mark D. Plumbley. "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network." *International Conference on Latent Variable Analysis and Signal Separation*. Springer International Publishing, 2015.

[23] Boersma, Paulus Petrus Gerardus. "Praat, a system for doing phonetics by computer." *Glot international* 5 (2001).

[24] http://www.fon.hum.uva.nl/praat/

[25] Itti, Laurent, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." *IEEE Transactions on pattern analysis and machine intelligence* 20.11 (1998): 1254-1259.

[26] Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *ICML*. Vol. 14. 2015.

[27] Vincent, Pascal, et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion." *Journal of Machine Learning Research* 11.Dec (2010): 3371-3408.