DOI: 10.17469/O2106AISV000023

ENRICO ZOVATO, VITO QUINCI, PAOLO MAIRANO

Modelling Sentiment Analysis scores and acoustic features of emotional speech with neural networks: A pilot study

Abundant literature has shown that emotional speech is characterized by various acoustic cues. However, most studies focused on sentences produced by actors, disregarding more naturally produced speech due to the difficulty in finding suitable emotional data. In our previous work we had performed an analysis of audiobook data in order to see if sentiment analysis could be of help in selecting emotional sentences from read speech. A regression analysis with Linear Mixed Models had revealed small effects, and the power of the models was low. We propose here an analysis with a neural network classifier predicting sentiment on the basis of acoustic cues, given the success of such models in the speech literature. However, the accuracy of the output was merely +0.13 above chance levels, suggesting that the different components used to express emotions (acoustic and lexical) tend to be complementary rather than additive, at least in audiobooks.

Key words: emotional speech, sentiment analysis, prosody, voice quality.

1. *Introduction*

The acoustic cues of emotional speech have been extensively described in the literature (Banse, Scherer, 1996; Burkhardt, Sendlmeier, 2000; Schröder, Cowie, Douglas-Cowie, Westerdijk & Gielen, 2001). However, most studies are affected by methodological issues, such as the use of acted speech, often produced by actors or other professional speakers. In this study, we contribute to the study of emotions on "non-explicitly-elicited" speech by investigating the correlation between Sentiment Analysis (SA) scores and acoustic cues of emotions in English audiobooks. In particular, we take in consideration a set of acoustic cues which have been associated to emotional speech in the literature, and investigate their relation with SA scores extracted from text. In fact, SA techniques apply to written text, but can of course be used with speech transcripts. Previous studies in this topic are largely missing, and the few studies that explored this topic found weak correlations between acoustic and lexical cues of emotions: Charfuelan and Schröder (2012) analyzed data of one audiobook and found mild but significant correlations. Mairano, Zovato and Quinci (2018) analyzed data from 251 audiobooks, extending the scope of Charfuelan and Schröder's experiment, and found significant effects which explained only small amounts of variance. In this contribution, we re-analyze the data used in our previous study (Mairano et al., 2018) using a neural network clas-

sifier. Neural networks are a powerful statistical tool that has recently been applied in many domains of the speech literature. Given the success of this technique in making accurate predictions of complex social phenomena, and in accounting for non-linear relationships among variables, we hope that their use will improve the accuracy and the power of our models.

For this work, we used the same data as Mairano et al. (2018), i.e. the *LibriSpeech* corpus, more precisely the subset containing data suitable for ASR models training. This set is called train_clean_100 and is composed of 100 hours of speech (28539 sentences) in English read by professional and non-professional speakers, male and female. Scripts are extracted from 251 books.

2. *Sentiment Analysis*

Most SA systems work at sentence or document level and are based on lexicons. In our framework, SA scores were extracted at sentence level by means of open source tools, namely Vader (Gilbert, Hutto, 2014) and SentiWordNet (Baccianella, Esuli & Sebastiani, 2012). Vader provides three values:

1. an index of positive polarity, in the range 0-1,
2. an index of negative polarity, in the range of 0-1,
3. a composite index that takes into account the previous values, in the range [1,1].

These scores are extracted by means of contextual rules and of the Vader Lexicon, which allows to assign a score to each word in the sentence. For example:

What a wonderful day, full of happiness and joy!
0 0 +4 0 0 0 +3 0 +3 → **10**

After being sacked, he felt horribly depressed.
0 0 0 0 0 -3 -2 → **-5**

One of the authors manually annotated a subset of 1000 sentences of the corpus as positive, negative or neutral, and compared the manual classification with Vader scores. By considering scores < 0.2 as negative, -0.2 < score +0.2 as neutral, and score > +0.2 as positive, the agreement between the manual classification and Vader scores was 72%.

Here are some examples of SA scores associated with sentences available in the *LibriSpeech* scripts:

Table 1 - *Examples of sentence SA scores*

King of glory king of peace hear the song and see the star welcome be thou heavenly king	+0.93
The juniors forever hurrah fans hurrah our team is a winner	+0.90
Now again I feel that bliss to love one's neighbors.	+0.84
Since she too is a lady whom I love most tenderly	+0.80
The most peaceful and lovely thing he had ever seen	+0.82
That no doubt was why she hated him	-0.84
But was defeated in a fierce battle on the banks of the Daban canal	-0.82
How they had cruelly robbed and murdered poor people	-0.91
How fearful and dizzy tis to cast one's eyes so low I'll look no more	-0.84
He was so enraged at this that he again began war	-0.81

3. *Acoustic Analysis*

As described in Mairano et al. (2018), we extracted acoustic features at sentence and word level. Most of the studies in this field focus on sentence-level analyses, however it is likely that the strong semantic emotional connotation of some words can be reflected by certain acoustic cues. Word level acoustic features were calculated isolating the portion of signal around the stressed vowel. This could be done because the whole dataset was forced-aligned at word, syllable and phoneme level. We analysed all words in the sentence except function words.

Audio data were segmented and phonetically aligned. In fact, we automatically transcribed all the text data with a General American English grapheme to phoneme tool. A speaker-independent HMM model was used to align the waveform data to their phonetic labels and; eventually, the output was converted into *Praat*'s textgrid format (Boersma, Weenink, 2018) for the successive analysis and processing. In summary, the acoustic analysis consisted of:

– Grapheme to Phoneme conversion;
– Forced alignment;
– Extraction of acoustic parameters;
– Data pruning.

Feature representation plays an important role in developing emotion related applications. Previous studies (Banse, Scherer, 1996; Burkhardt, Sendlmeier, 2000; Schröder, Cowie, Douglas-Cowie, Westerdijk & Gielen, 2001), which extensively analyzed acoustic features, led us to select and calculate the following pitch features (*f0*) with a custom Praat script:

– F0 mean (in semitones);
– F0 stdev, standard deviation of F0 in semitones;
– F0 range (0.05 – 0.95);
– F0 max (0.95);
– F0 min (0.05).

F0 data were extracted with the autocorrelation method available in Praat and was executed in two steps. During the first iteration, fixed values of minimum and

maximum F0 were used (75-400 Hz). The results were used to determine the inter quartile range (IQR). In the second step, the same analysis was run using a range between -25% and + 50% of IQR. Apart from F0 range, F0 min and max were included to explicitly define the level of F0.

Additionally, the following spectral features (*spec*) were extracted:

– Shimmer, as the perturbation in the amplitude of consecutive periods in voiced sounds;
– Jitter, as the perturbation in the duration of consecutive periods in voiced sounds;
– HAM, Hammarberg Index, as the difference of peak energy in the bands 0-2 kHz and 2-5 kHz (Hammarberg, Fritzell, Gauffin, Sundberg & Wedin, 1980);
– Do1000, as the linear declination of energy above 1000 Hz;
– Pe1000, relative energy of frequencies above 1000 Hz with respect to energy below 1000 Hz (Sherer, 1989; Drioli, Tisato, Cosi & Tesser, 2003);
– HNR, Harmonic to Noise Ratio;
– NHR, Noise to Harmonic ratio.

Beyond the features mentioned above, further duration parameters (*dur*) were extracted at sentence level, namely:

– DUR, total duration of the sentence from the beginning of the first phoneme to the end of the last one;
– SR, speech rate. The number of phonemes divided by the total length of the sentence, including the pauses;
– AR, articulation rate. The number of phonemes divided by the total length of the sentence, excluding the pauses;
– PSR, pause/speech ratio.

All the extracted parameters were converted into z-scores for each speaker (i.e. normalized with respect to the average and standard deviation values of the corresponding speaker). Different combinations of the abovementioned set of features (*f0*, *dur*, *spec*) were used in later experiments in order to isolate their effect on the classification models.

We discarded utterances whose duration was below 3 seconds. Moreover, all the acoustic parameters exceeding 2.5 times the standard deviation were excluded, as likely detection errors.

4. *Modeling SA scores and acoustic features*

We analyzed the acoustic and SA scores with two techniques. The initial analysis (reported in Mairano et al., 2018), was performed with linear mixed effects regression, using R and the *lme4* library (Bates, Maechler, Bolker & Walker, 2014) and revealed a weak linear relation between SA scores and acoustic features of emotional speech. Some effects were indeed significant, but the effect size was often extremely small. This was reflected in the power of the models, which was also disappointingly low. In this contribution, we present a follow-up experiment where we trained a neural network SA classifier, whose input was composed of the acoustic features.

The output of the model was tested on a specific test-set, which had not been used for training.

4.1 Sentiment Analysis Neural Net Classifier

Neural networks are powerful parallel processors that can acquire knowledge through observed data. They can be used both for regression and classification tasks (LeCun, Boser, Denker, Henderson, Howard, Hubbard & Jackel, 1990). Their effectiveness lies in the capability to model complex non-linear transformations between input and target data. The deepness in layers helps abstract the input layer information at increasing levels of complexity.

In our experiment we opted for supervised learning applied to multi-layer feed-forward networks, given the nature of input and output data. Training the network means determining the inner weights and biases that tend to reduce the error between predicted and observed data. The training is executed by using the backpropagation algorithm, which is composed of two phases: the forward phase, in which the weights are fixed, and the input is propagated through the network to calculate the error between the predicted and actual output; and the backpropagation phase, where the error is propagated through the network in the backward direction and weights are adjusted to minimize the error.

In our experiment we wanted to predict SA classes from input acoustic features. In other words, our goal was to find a transformation function from the acoustic level to the valence level (text based). We chose a relatively simple architecture, opting for a multi-layer feed-forward network, that is a graph in which all nodes of a layer are fully connected to the nodes of the preceding/following layer, and there are no recurrences. The transformation function can be written as:

$$h(x) = g(\sum_i w_i \, h_i(x) + b_i) \quad (1)$$

where g is the output activation function, i is the layer index, are the layer functions, are the weights, and are the biases.

The network was composed of three feed-forward hidden layers and we used Rectified Linear Units (ReLU) as activation functions. On top of them, we stacked a softmax layer, which simply converts the output activation values to class probabilities which sum to 1. The softmax function is:

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}} \quad (2)$$

The input of the network was composed of the acoustic features described in section 3. The targets were the SA categories, according to the compound sentiment scores classification:

- negative (Neg)
- neutral (Neu)
- positive (Pos)

Our model was tested on a portion of data excluded from training, in order to assess the accuracy of the model and its generalization capability.

Two classifiers were trained, the first with sentence-level features as input (SNT), the second with word-level features, extracted from lexically stressed vowels of content words (WRD).

The features used in WRD models were: DUR, HNR, NHR, F0 min, F0 max, F0 mean, F0 stdev, jitter, shimmer, Do1000, Pe1000, HAMM. For SNT models, AR and SR were also included.

A further distinction was made about the number of target categories: we considered three-class models (Neg, Neu, Pos), as well as binary classification with just Pos and Neg classes.

The network weights were trained on the whole train_clean_100 set of LibriSpeech data. During training, an early stopping criterion was adopted, on the basis of loss values of the validation set (30%), and on 10 consecutive non-improving epochs. Adam optimizer (Kingma, Ba, 2015) was used and the loss function was based on categorical cross-entropy. Input data were normalized to zero mean and unit variance. Data sampling was also applied in order to have a balanced number of occurrences among the three/two SA categories. A subset of the data (10%), which was not used for training, was later used as a test set for evaluation.

4.2 Results

We initially run an experiment in which all the acoustic features were used as input to the network model. We evaluated the models on the test set and results indicated low levels of accuracy levels, barely above chance level. This happens for both two- and three-class models. To some extent, these results confirm the outcomes of Mairano et al. (2018), indicating that correlations between the textual (SA) and acoustic levels are weak, at least in audiobook data. Accuracy, precision and recall values are reported in Table 2.

Table 2 - *Accuracy, precision and recall scores of SA by model*

	Accuracy	**Precision**	**Recall**
WRD-2	0.561	0.555	0.619
WRD-3	0.400	0.402	0.400
SNT-2	**0.638**	0.633	0.690
SNT-3	**0.462**	0.463	0.463

SNT-2 and SNT-3 models provide better accuracy than WRD-2 and WRD-3 models respectively. This indicates that sentiment scores are more predictable in wider contexts, or on global rather than local parameters. This can be also observed in the confusion matrices, that are a useful technique which describes the performance of a classification model. The rows and columns of these matrices represent respectively the predicted and true classes. The diagonal of a confusion matrix indicates the true positives (i.e. cases in which the predicted class matches the true class). As it can be observed in Figures 1, 2, 3 and 4, SENT models have a more "diagonal"

pattern, despite the significant classification errors. In these figures the darker the color in the diagonal, the higher is the match between predicted and actual labels. Hence, we consider this model as the best performing (even though the accuracy score might point otherwise).

Figure 1 - WRD-3 *evaluation confusion matrix*

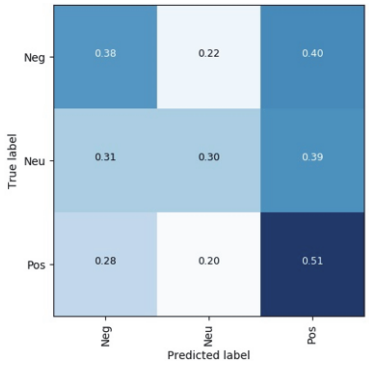


Figure 2 - SNT-3 *evaluation confusion matrix*

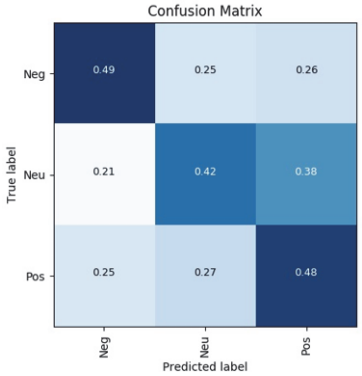


Figure 3 - WRD-2 *evaluation confusion matrix*

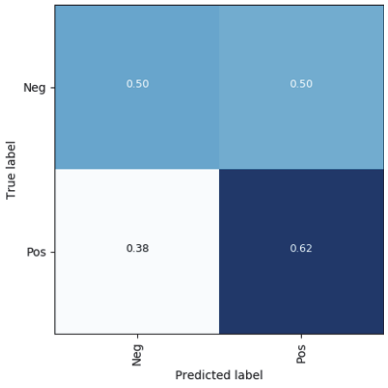
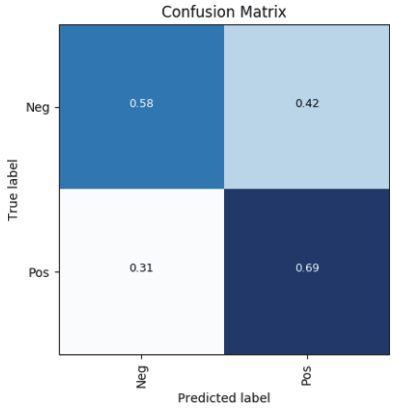


Figure 4 - SNT-2 *evaluation confusion matrix*



A second experiment analyzed different sets of input features applied to the SNT-3 models. In fact, according to the previous experiment, sentence models provided better results. Moreover, the 3-class model, despite its slightly lower accuracy than the 2-class model, (SNT-2 and SNT-3 are 0.138 and 0.129 above chance level, respectively), fulfills a more extended task, in line with more refined data classification. In this context we trained SNT-3 models whose input layer was composed of duration features (*dur*), pitch features (*f0*), duration+pitch features (*pros*) and spectral features (*spec*), as reported in Table 3. The motivation behind this experiment was to analyze the relative contributions of the features to model accuracy.

We tested the models on the same test set as the previous experiment. Results indicate that none of these variants reaches the accuracy of the "full set" model, meaning that all features bring relevant information for the prediction. The feature set that is closer in accuracy to the "full set" is the *pros* one: this result seems to confirm some results observed in linear mixed effects models in Mairano et al. (2018), according to which, prosodic features provided the higher correlation rates. On the other hand, duration and spectral features alone produce the worst results in terms of accuracy.

Table 3 - *Accuracy, precision and recall scores of SA SNT-3 models, trained on subsets of the acoustic features*

SNT-3	Accuracy	Precision	Recall
pros	**0.4397**	0.4407	0.4402
spec	0.3706	0.3711	0.3725
dur	0.3950	0.4288	0.3944
f0	0.4168	0.4167	0.4170

Table 4 - *Feature groups*

dur	DUR, SR, AR, PSR
f0	F0 mean, F0 stdev, F0 range, F0 max, F0 min
spec	Shimmer, Jitter, HAM, Do1000, Pe1000, HNR, NHR

5. *Discussion and Conclusions*

In this study we have analyzed audiobook data trying to find correlations between textual polarity, as provided by SA tools, and acoustic features extracted from the corresponding speech data. The results of our experiments have shown that SA classification from acoustic data is challenging. We have used neural network models which can learn complex transformations between input features and target categories. The underlying reason was to determine whether this modeling technique can provide better accuracy than other techniques used in past experiments (cfr. Mairano et al., 2018). Results, however, confirm the existence of a relation between pitch parameters and polarity text SA scores. Nonetheless, such relation is not strong enough to allow for a reliable prediction of SA categories. The prediction accuracy is merely 0.13 points above chance level in the best case (i.e. three-class sentence models). The correlation between lexical and acoustic cues of emotion seems thus to be limited in audiobooks, possibly because of a reduced involvement on the part of the speaker with respect to the content of the utterance. This seems to suggest that textual and acoustic cues are mostly complementary, rather than cumulative, in this type of data. If so, SA may not be well suited for the selection of emotional speech from audiobooks, or more generally for the study of emotions in speech.

The choice of audiobooks is dictated by the availability of many data resources in which a match between the textual content and an appropriate speaking style should occur. The problem, as we have shown in our experiments, is that either this kind of speech is not too much emotionally characterized or the SA tools we used are not adequate for capturing emotional content in the audiobook texts. Using specific databases of elicited emotional speech produced by actors or professional speakers would shed some light on this aspect. The risk is however that these data are too prototypical or overacted, sounding too distant from natural speech.

In future work we will try to extend the analysis to more data, taking into account audiobooks read by professional speakers. Another aspect to further investigate is the effect due to the sequence of words and their SA scores, as well as the sequence of acoustic features. In this regard, more complex neural network architectures will be considered, such as multi-layer recurrent networks with long-short term memory cells.

Bibliography

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC*, La Valetta, 17-23 May 2010, 2200-2204.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67, 1, 1-48.

Boersma P., Weenink D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

Burkhardt, F., Sendlmeier, W.F. (2000) Verification of acoustical correlates of emotional speech using formant-synthesis. In *SpeechEmotion-2000*, 151-156.

Charfuelan, M., Schröder, M. (2012). Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives. In *Proceedings of 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals ES3*, Istanbul, 26 May 2012, 99-103.

Drioli, C., Tisato, G., Cosi, P. & Tesser, F. (2003). Emotions and voice quality: experiments with sinusoidal modeling. In *Proceedings of Workshop the Voice Quality: Functions Analysis and Synthesis (VOQUAL)*, Geneva, 27-29 August 2003, 127-132.

Gilbert C.J., Hutto E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI, 2-4 June, 2014.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities In *Acta Otolaryngologica*, 90, 441-451.

Kingma, D.P., Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR*, San Diego, CA, 7-9 May 2015.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. & Jackel, L.D. (1990). Handwritten digit recognition with a back-propagation network. Advances. In *Neural Information Processing Systems*, 2, 396-404.

Mairano, P., Zovato, E. & Quinci, V. (2018). La sentiment analysis come strumento di studio del parlato emozionale?. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it*, Turin, 10-12 December 2018.

Scherer, K.R. (1989). Vocal correlates of emotion. In Manstead, A., Wagner, H. (Eds.), *Handbook of psychophysiology: Emotion and social behavior*. London: Wiley, 165-197.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M. & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings EUROSPEECH 2001 – Seventh European Conference on Speech Communication and Technology*, Aalborg, 3-7 September 2001, 87-90.