Do sentiment analysis scores correlate with acoustic features of emotional speech?

Paolo Mairano[1], Enrico Zovato[2], Vito Quinci[2]

[1]University of Lille (France), [2]University of Turin (Italy)

Theoretic background and motivation

Emotional and affective information can be conveyed at different levels [1]: lexical (intensifiers, modals, hedges, etc.), syntactic (e.g. relative clauses to comment on actions and behaviors), paralinguistic (facial expressions, gestures), and by voice. The mainstream frameworks for the analysis of emotional and affective voice characteristics are based either on categories such as anger, excitement, disgust, fear, relief, sadness (e.g. [2]), or on dimensions (e.g. [3]). The latter tend to define emotions as coordinates in a multidimensional space, each dimension representing a property of an emotional state. For instance, the framework proposed by [3] postulates two dimensions of emotions, namely valence (positive vs. negative) and arousal (high vs. low activation): a sad state can be defined as having negative valence and low activation, while a euphoric state can be defined as having positive valence and high activation. The classification of emotional states has proved difficult within categorical as well as dimensional frameworks, and the situation is further complicated by the interfering role of other linguistic levels (lexical, syntactic), as mentioned below: it is not yet clear whether lexical and syntactic features can be considered as ancillary or complementary to voice characteristics for the expression of emotions, and whether the relation between these two levels of analysis can change according to the style or the context of communication.

Despite such intrinsic problems, many studies have focused on emotional speech with the aim of finding acoustic correlates of affective states. For example, it has been found that limited pitch variations characterize affective states with negative valence, while higher pitch and higher pitch range generally characterize affective states with positive valence [4]. Beyond pitch, other prosodic features seem to play an important role in conveying emotions, and various other acoustic and perceptive correlates of emotions have been suggested in other studies (|6], [7], [8]). However, one of the problems affecting such studies is the availability of reliable data. Given the difficulty of obtaining controlled emotional data elicited in an ecological context, most studies make use of acted speech, with the clear drawback that the resulting emotional speech tends to be prototypical or overacted [9] and does not necessarily correspond to natural realizations. In this contribution, we investigate the correlation between sentiment analysis metrics and acoustic characteristics of speech, as measured on audiobooks. A similar analysis investigating the correlation between sentiment analysis scores and acoustic features has been attempted by [10], but only on data from one speaker and one audiobook. We extend the analysis to 251 audiobooks in the hope to serve a practical purpose (the investigation of whether sentiment analysis can be of help in gathering ecological emotional speech for analysis) and to contribute to the theoretic discussion about the roles of voice cues vs lexicon in expressing emotions.

Data and methodology

In order to study the correlation of sentiment analysis scores and acoustic features of emotional speech, we used audiobooks, as per [10]. Audiobook recordings came from the *LibriSpeech* ASR corpus, more precisely the train-clean-100 section, containing 100 hours of clean speech from 251 audiobooks read by different speakers. All the material was transcribed phonetically with a TTS front-end component, following transcription conventions of General American. Phonetic transcriptions were then forced-aligned to the acoustic signal enabling silence detection, and finally converted to *TextGrid* format for analysis with *Praat*.

Sentiment analysis scores were extracted from text for each sentence using *Vader* and *SentiWordNet*, both available within Python's NLTK library. We extracted acoustic features at sentence level and at word level in *Praat*: mean F0 (in semitones), pitch stdev (in semitones), pitch range (0.05 to 0.95 quartiles), pitch max (0.95 quartile), pitch min (0.05 quartile), shimmer, jitter, Hammarberg index (HAM, difference between max energy in the 0-2 kHz and 2-5kHz bands, [11]), Do1000 (drop off spectral energy above 1000 Hz), Pe1000 (relative energy in the frequencies above 1000 Hz versus energy below 1000 Hz, [12], [13]). For the sentence-level analysis, we also extracted the total duration in ms from first to last phoneme (DUR), speech rate (SR), articulation rate (AR), and pause/speech ratio (PSR). All the acoustic parameters were transformed to z-scores by speaker, in the attempt to normalize individual differences.

Results and discussion

The data were entered in linear mixed-effects models evaluating the relation between sentiment analysis scores and acoustic parameters. We then built separate models for negative (*Vader* score < 0) vs positive (*Vader* score > 0) sentences. Models for sentences with positive Vader values showed that *Vader* positivity score was a significant predictor not only

for pitch parameters, but also for rhythmic and spectral parameters. Similarly, models for sentences with negative Vader values showed that *Vader* negativity value was a significant predictor not only for pitch parameters, but also for AR and shimmer. These results seem to suggest that, while pitch parameters show a (modest) linear correlation with valence, rhythmic and spectral parameters correlate with arousal.

The acoustic features have then been used to train a neural network classification model, whose targets were sentiment analysis categories. The goal was to verify whether non-linearities involving acoustic features can contribute to some extent to the prediction of sentiment classes. Two classifiers were trained, one based on sentence-level features, the other on word-level features extracted on the stressed vowel. We considered three-classes models (Neg, Neu, Pos), as well as binary classification with just Pos and Neg classes. Multilayer perceptrons were trained for this experiment on the whole set of *LibriSpeech* data. They were composed of three hidden layers with ReLU activation functions and an output softmax layer. An early stopping criterion was adopted during training, on the basis of loss values of the validation set (30%), and on 10 consecutive not-improving epochs. Adam optimizer 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 such a way to have a balanced number of occurrences among the three/two categories. A subset of data (10%) not used for training were then used for evaluation. Results in Table 1 show that accuracy is little above chance level in two and three classes models.

Accuracy	**Word**	**Sentence**
3-classes	0.400 (0.402, 0.400)	0.460 (0.463, 0.461)
2-classes	0.561 (0.555, 0.619)	0.638 (0.633, 0.690)

Table 1: *Classification accuracy (precision and recall) results of NN sentiment predictors. The sentence-level gain is 0.12 with respect to chance level (which is 0.33 for three classes, 0.5 for two classes).*

These results suggest that lexical and acoustic cues of emotion do not necessarily go hand-in-hand, so that speakers/listeners may rely on either in order to realize or perceive emotions in speech. This may be especially true for read speech where the reader is not personally involved in the content, as for audiobooks. If so, the aptness of sentiment analysis for the study of emotional speech may be limited, since relying on such metrics would mean disregarding any acoustic cues of emotion that do not co-occur with lexical cues. However, we acknowledge a certain number of limitations of this study. Firstly, despite the undeniable advantages of a fully automatic approach for the transcription, annotation and analysis, the procedure is likely to contribute to a certain amount of noise in the data. This potentially has the effect of reducing the observable relations among the variables studied. Secondly, the lack of punctuation in *LibriSpeech* corpus sentences makes it impossible to differentiate narrative chunks from reported speech by characters, where more pronounced emotional features can be expected. Finally, rule-based open-source tools for sentiment analysis certainly provide a convenient solution, but more accurate sentiment analysis algorithms may yield different results.

References

[1] J. Reilly and L. Seibert, "Language and emotion," in *Handbook of affective sciences*, R.J. Davidson, K.R. Scherer and H.H. Goldsmith, Eds., OUP, 2003, pp. 535–559.

[2] P. Ekman, "Basic Emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and T. Power, Eds., 39.6, London (UK), John Wiley & Sons, 2000, pp. 45-60.

[3] J.A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no.6, 1161-1178, 1980.

[4] F. Burkhardt, W.F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *SpeechEmotion-2000*, pp. 151-156, 2000

[5] T. Johnstone and K.R. Scherer, "Vocal communication of emotion," in *Handbook of emotions 2*, M. Lewis and J. Haviland, Eds., London-New York: The Guildford Press, pp. 220–235, 2000.

[6] R. Banse and K.R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, no. 3, 614–636, 1996.

[7] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *EUROSPEECH 2001 – Seventh European Conference on Speech Communication and Technology*, September 3-7, Aalborg, Denmark, Proceedings, 2001.

[8] N. Audibert, V. Aubergé and A. Rilliard, "The prosodic dimensions of emotion in speech: the relative weights of parameters," *Ninth European Conference on Speech Communication and Technology*, 4-8 September, Lisbon, Portugal, 2005.

[9] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, 227-256, 2003.

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

[11] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolaryngologica*, vol. 90, 441-451, 1980.

[12] K.R. Scherer, "Vocal correlates of emotion," in *Handbook of psychophysiology: Emotion and social behavior*, A. Manstead and H. Wagner, Eds., pp. 165-197, London: Wiley, 1989.

[13] C. Drioli, G. Tisato, P. Cosi, F. Tesser, "Emotions and voice quality: experiments with sinusoidal modeling," in *Voice Quality: Functions Analysis and Synthesis (VOQUAL) Workshop*, 27-29 August, Geneva, Switzerland, Proceedings, 2003.