# Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives

**Marcela Charfuelan, Marc Schröder**

Language Technology Laboratory, DFKI GmbH
Berlin D-10559 and Saarbrücken D-66123, Germany
{marcela.charfuelan,marc.schroeder}@dfki.de

**Abstract**

We investigate possible correlations between sentiment analysis scores obtained for sentences of Mark Twain's novel "The Adventures of Tom Sawyer" and acoustic features extracted from the same sentences in the corresponding audiobook. We have found that scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative, in particular more correlated with average energy and mean fundamental frequency (F0). We have designed an experiment intended to predict the levels of acoustic expressivity in arbitrary text using sentiment analysis scores and the number of words in the text.

## 1. Introduction

In this paper we investigate possible correlations between sentiment analysis scores obtained for sentences of Mark Twain's novel "The Adventures of Tom Sawyer" and acoustic features extracted from the same sentences in the corresponding audiobook. In the audiobook a single speaker reads the whole novel, the narration is lively and expressive and the speaker impersonates or performs several characters apart from the narrator himself.

From a theoretical point of view, narratives have been studied as a context for the integration of language and emotion. According to (Reilly and Seibert, 2003) and the references in this work, evaluative information in narratives can be conveyed/packaged in several ways: "lexically", for example, using intensifiers, modals or hedges to reflect speaker attitude; "syntactically" as in relative clauses, which commonly function as asides to comment on a person's behaviour/character; and "paralinguistically", by emotional facial expression, gesture and affective prosody that can effectively convey narrator attitude or reflect the inferred emotions of a character.

Due to the lively character of narration in audiobooks, these have been recently used in several studies related to clustering of expressive speech styles (Székely et al., 2011), expressivenes of speech (Wang et al., 2006) or automatic selection of diverse speech corpora for improving automatic speech synthesis (Braunschweiler et al., 2011a). Audiobooks might help to tackle some of the nowadays key problems on speech synthesis technology: unlabelled prosodic and voice quality variations; expressive speech; large corpora of non-studio-quality speech (Blizzard Challenge, 2012). At the same time audiobooks might also contribute to simplify some of the most difficult problems to progress with synthesis from social signalling corpora: lack of phonetic coverage, lack of single-user speech, and lack of textual transcriptions.

In this paper first we describe the data analysed in Section 2., then in Sections 3. and 4. we describe the sentiment scores obtained for sentences in the book and the acoustic features extracted from the corresponding audio data. In Section 5. we describe two experiments intended to investigate the possible correlation of the previous scores and features and the possibility of using sentiment scores from arbitrary text to predict an acoustic level of "expressivity". Preliminary results and future work are presented in Section 6.

## 2. Data

The data analysed is the audiobook "The adventures of Tom Sawyer" available at LibriVox (LibriVox, 2012) and its associated text available in Project Gutenberg (Project Gutenberg, 2012). The audiobook has been split into prosodic phrase level chunks, which corresponds to the sentences analysed in this work. The sentence segmentation and orthographic text alignment of the audiobooks has been performed using an automatic sentence alignment method - LightlySupervised - described in (Braunschweiler et al., 2011b). The number of sentences analysed is 5119 corresponding to 17 chapters and approximately 6.6 hours of recordings at 44100 Hz. The books were read by John Greeman, an American English narrator.

## 3. Sentiment scores

The sentiment scores were obtained in two steps. First, summary statistical information about individual words was extracted using the data and methods of (Potts and Schwarz, 2010) and (Potts, 2011a). Second, to combine these word-level scores effectively in order to make predictions about full sentences, a maximum entropy classifier was trained on a large, diverse collection of texts from social media sources. The reader is referred to these publications for more details about the system as well as (Potts, 2011b) for data and available resources. In the following we summarise the sentiment scores used in this study:

- Scores derived from IMDB reviews using machine learning techniques (Bo et al., 2002):
  - ImdbEmphasis: a sentiment score for emphasis vs. attenuating
  - ImdbPolarity: a sentiment score for positive vs. negative
- OpinionLexicon: sentiment scores by lexicon lookup using Bing Liu's lexicon, which is a list of positive

and negative opinion words or sentiment words for English (around 6800 words) that has been compiled over many years (Liu, 2011).

- SentiWordnet (Wordnet entries with added sentiment scores) negative and positive value:
  - SentiWordNetNeg
  - SentiWordNetPos
- Scores derived from the Experience Project: this project is a social networking website that allows users to share stories about their own personal experiences, users write typically very emotional stories about themselves, and readers can then chose from among five reaction categories to the story (Potts, 2011b). Data from this project has been used to derive the following reaction scores:
  - Hugs: Sympathy reader reaction score
  - Rock: Positive-exclamative reader reaction score.
  - Teehee: Amused/light-hearted reader reaction score.
  - Understand: Solidarity reader reaction score.
  - Wow: Negative-exclamative reader reaction score.
- Predicted negative (Neg) and positive (Pos) probability derived by training a model with the previous scores:
  - Neg, Pos
  - Polar: calculated as Pos-Neg, this is a kind of predicted polarization score, examples of very positive and very negative polarity scores are presented in Table 1.

| Text | Polar |
|---|---|
| Well, goodness gracious! | 1.00 |
| Luck! | 1.00 |
| I love thee well! | 1.00 |
| Glory was sufficient. | 0.99 |
| Tom's astonishment was boundless! | 0.99 |
| Good! | 0.99 |
| ... | |
| Kill? | -1.00 |
| It's awful. | -1.00 |
| Hateful, hateful, hateful! | -1.00 |
| Crash! | -1.00 |
| Bother! | -1.00 |
| It's that dreadful murder. | -1.00 |

Table 1: Text examples of very positive and very negative polarity scores.

## 4. Acoustic features

We have extracted well known acoustic correlates of emotional speech: mainly prosody or fundamental frequency (F0) related features, some intonation related measures (F0 contour measures) and voicing strengths features, that have been used to model and improve excitation in vocoded speech. The following features and measures have been calculated:

- F0 and F0 statistics, mean, maximum, minimum and range. F0 values were extracted with the snack tool (Sjölander, 2012).
- Duration in seconds per sentence.
- Average energy, calculated as the short term energy ($\sum s^2$) averaged by the duration of the sentence in seconds.
- Number of voiced frames, number of unvoiced frames and voicing rate calculated as the number of voiced frames per time unit.
- F0 contours, as in (Busso et al., 2009) we have extracted slope (a1), curvature (b2) and inflexion (c3); these measures are estimated by fitting a first-, second- and third-order polynomial to the voiced F0 values extracted from each sentence:

$$y = a_1 * x + a_0 \quad (1)$$

$$y = b_2 * x^2 + b_1 * x + b_0 \quad (2)$$

$$y = c_3 * x^3 + c_2 * x^2 + c_1 * x + c_0 \quad (3)$$

- Voicing strengths estimated with peak normalised cross correlation of the input signal (Chu, 2003). The correlation coefficient for a delay $t$ is defined by :

$$c_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+1)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}} \quad (4)$$

  Five bandpass voicing strengths are calculated, that is, the input signal is filtered into five frequency bands; mean statistics of this measure are extracted.

## 5. Experiments

### 5.1. Correlation analysis

Pairwise correlation between the previously described sentiment scores and acoustic features was performed. We have found correlations mainly between average energy and mean F0 and sentiment scores derived from IMDB reviews and reader reaction scores. Table 2 shows the higher correlation values between these scores and features. The correlation with other sentiment features was very low, in particular no correlation at all was found between F0 contour features and sentiment scores. These results also show that the sentiment scores that come from lexicons are not correlated at all with acoustic features, whereas scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative.

| | Acoustic features | |
|---|---|---|
| Sentiment scores | Energy | mean_F0 |
| ImdbEmphasis | 0.51 | 0.38 |
| ImdbPolarity | -0.33 | -0.31 |
| Teehee | 0.29 | 0.13 |
| Wow | -0.17 | -0.30 |
| Polar | -0.13 | -0.14 |

Table 2: Pairwise correlation between sentiment scores and acoustic features.

### 5.2. Predicting "expressivity"

In a further experiment we investigate if we can predict some measure of "expressivity" just on the basis of sentiment scores. Our measure of expressivity is the first principal component value (PC1) after a principal component analysis (PCA) of all the acoustic features extracted from the data. A PC1 value per sentence was calculated, and we have empirically found that positive values of PC1 most of the time correspond to sentences of the narrator in a more or less neutral voice, and negative values most of the time correspond to expressive sentences where the speaker impersonates one of the characters in the book (childish voice, women voice. etc.). To corroborate this, we have manually annotated the first two chapters of the book according to narrator and the characters the speaker performs. Figure 1 shows the variation of mean F0, ImdbEmphasis and PC1 per sentence in chapter 01, for the narrator and other impersonated characters. In this Figure we can also observe that the values for "other" characters present higher excursion than for the "narrator".

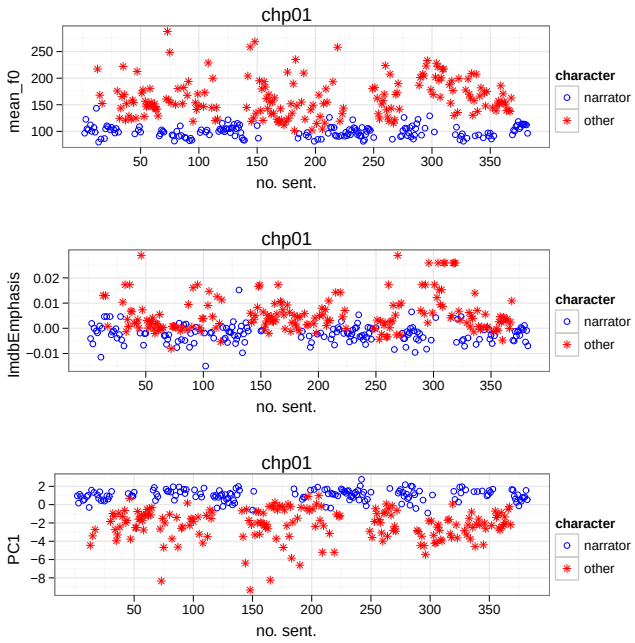


Figure 1: Mean F0, ImdbEmphasis scores and PC1 values for the sentences in chapter 01 of "The Adventures of Tom Sawyer".

Multiple linear regression (MLR) of sentiment scores (plus number-of-words) was used to train a prediction model of the acoustic PC1 feature; sequential floating forward selection (SFFS) was used to find the best sentiment score predictors. Statistical analysis, MLR and SFFS, were performed with R (R Development Core Team, 2009). We have found that the model fits well the training data. Figure 2 shows in blue the PC1 values obtained per sentence for chapter 02 of the book; the predicted values are indicated in red and the prediction error in black. Averaging the results obtained for every chapter, we have found that PC1 is predicted with a prediction error of 1.21 when using just sentiment features; the prediction error improves to 0.62 when using number-of-words in the sentence as another predictor feature.

To evaluate how well the model can predict a level of expressivity with unseen data, we used the annotated chapters 01 and 02 as test data and the rest of the data to train a model. For training a predictor model of PC1 we used all the acoustic features presented en Section 4.; the learnt parameters after the SFFS multiple linear regression are:

$$\begin{aligned} PC1 = &-1.64 + 0.12 \times num\_words\_sentence \\ &- 48.0 \times ImdbEmphasis + 11.3 \times ImdbPolarity \\ &+ 2.24 \times SentiWordNetNeg - 1.78 \times Teehee \\ &- 3.66 \times Understand - 1.17 \times OpinionLexicon \\ &+ 0.6 \times Hugs + 0.44 \times SentiWordNetPos \end{aligned} \quad (5)$$

Using this equation a PC1 value is predicted for the utterances of chapters 01 and 02, the value is further used to determine whether the utterance is character type "narrator" (predicted PC1 $>= 0$) or "other" (predicted PC1 $< 0$). Since we have character annotations of these two chapters we can compare the annotated character and the predicted one. The character prediction results for 345 utterances of chapter 01 and 271 utterances of chapter 02 are presented in Table 3. Examples of utterances predicted as "narrator" and "other" in chapter 01 are presented in Table 4.

| | Chapter 01 | | Chapter 02 | |
|---|---|---|---|---|
| Character | Narrator | Other | Narrator | Other |
| Narrator | 79.8 | 30.1 | 92.0 | 34.0 |
| Other | 20.2 | 69.9 | 8.0 | 66.0 |
| Diagonal | 73.3% | | 81.5% | |

Table 3: Character prediction for chapters 01 and 02 using number of word, sentiment scores and the learnt model in equation 5.

We can observe in Table 3 that the character types in chapter 02 were better predicted than in chapter 01. Two observations might explain why "expressivity" in chapter 01 was more difficult to predict: first, the PC1 values of chapter 01 present higher excursion than chapter 02 and second the sentences in chapter 01 are shorter in average than in chapter 02. Chapter 01 has 12.3 words in average per sentence (minimum 1 and maximum 80 words) and chap-

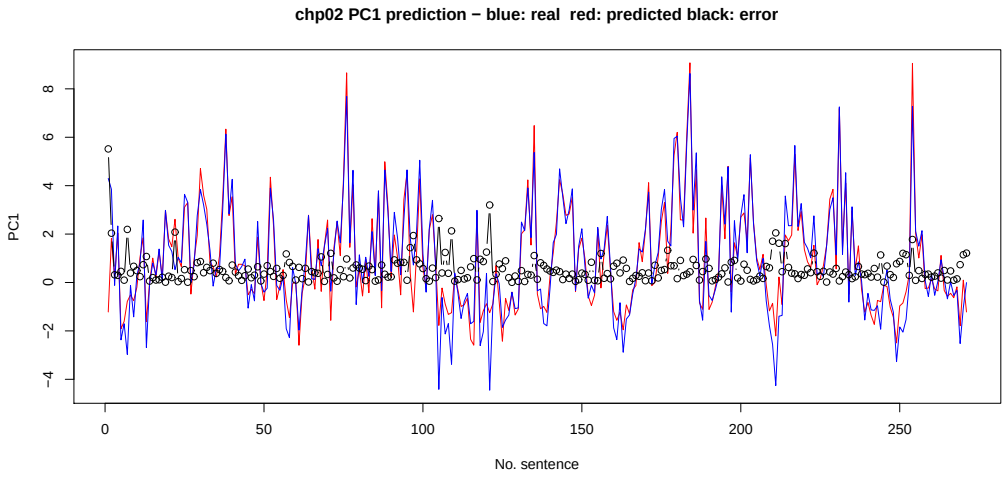


Figure 2: Prediction of PC1 using multiple linear regression of sentiment analysis scores and number of words in the sentence for chapter 02 of "The Adventures of Tom Sawyer".

ter 02 has 20.5 words in average per sentence (minimum 1 and maximum 93 words). These observations confirm that short sentences tend to be more expressive and therefore more unpredictable in terms of sentiment analysis (Mohammad, 2011). The sentences presented in Table 4, exemplify this difficulty, although from an acoustic point of view the model is able to capture quite well the style intended by the reader in the book. In fact auditively the sentences presented in this Table are quite different, which makes it possible to define and predict more than two expressive styles.

## 6. Conclusions

We have found that sentiment analysis scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative, in particular more correlated with average energy and mean F0. Scores derived from lexicon and Sentiwordnet are much less correlated with the acoustic features in the analysed data. It is interesting to notice that any of the F0 contour measures (intonation measures) correlate with sentiment scores, this observation probably is in line with the findings of (Busso et al., 2009) where it has been found that gross pitch statistics are more emotionally prominent than features describing the pitch shape.

We have designed an experiment intended to predict the levels of acoustic expressivity in arbitrary text using sentiment analysis features and the number of words in the text. We have found that the predictive model fits well the training data, and it is able to predict the style of unseen data, in particular the character style of utterances in two chapters of the book not used for training the model.

An immediate application of these results is in automatic speech synthesis. We have demonstrated that an style can be automatically derived from textual data and a trained model, so the next step is to use this information to select the expressive style with wich the text should be realised. Also, given the clear auditive differentiation of utterances along PC1 values we will consider to predict more than two styles defining various PC1 thresholds.

## 7. Acknowledgements

This work is supported by the EU project SSPNet (FP7/2007-2013). We would like to thank Christopher Potts for providing us with the sentiment analysis of the data and Holmer Hemsen for assistance on data annotation.

## 8. References

Blizzard Challenge. 2012. Blizzard Challenge 2012. http://www.synsig.org/index.php/Blizzard_Challenge_2012.

Pang Bo, Lee Lillian, and Vaithyanathan Shivakumar. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA.

N. Braunschweiler, , and S. Buchholz. 2011a. Automatic sentence selection from speech corpora including diverse speech for improved hmm-tts synthesis qualityt=/. In *Interspeech*, Makuhari, Chiba, Japan.

N. Braunschweiler, M.J.F. Gales, and S. Buchholz. 2011b. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Interspeech*, Makuhari, Chiba, Japan.

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4):582–596.

Wai C. Chu, 2003. *Mixed excitation linear prediction*, chapter 17, pages 454–485. Speech coding algorithms Foundations and Evolution of Standardized Coders. Wiley.

LibriVox. 2012. LibriVox acoustical liberation of books in the public domain. http://librivox.org/.

Bing Liu. 2011. Opinion mining, sentiment analysis, and opinion spam detection.

| Character | Predicted_PC1 | Text |
|---|---|---|
| narrator | 3.00 | Soon the free boys would come tripping along on all sorts of delicious expeditions, and they would make a world of fun of him for having to work -- the very thought of it burnt him like fire. |
| narrator | 2.24 | He then held a position at the gate for some time, daring the enemy to come outside, but the enemy only made faces at him through the window and declined. |
| narrator | 1.69 | If one moved, the other moved -- but only sidewise, in a circle; they kept face to face and eye to eye all the time. |
| narrator | 0.58 | So she lifted up her voice at an angle calculated for distance and shouted: |
| ... | | |
| narrator | 0.05 | Spare the rod and spile the child, as the Good Book says. |
| narrator | 0.04 | I reckon you're a kind of a singed cat, as the saying is -- better'n you look. |
| narrator | 0.01 | If you was to tackle this fence and anything was to happen to it -- " |
| other | -0.00 | Another pause, and more eying and sidling around each other. |
| other | -0.00 | Ben ranged up alongside of him. |
| other | -0.02 | He opened his jacket. |
| other | -0.04 | "Tom, it was middling warm in school, warn't it?" |
| other | -0.05 | At this dark and hopeless moment an inspiration burst upon him! |
| ... | | |
| other | -1.87 | "Nothing." |
| other | -1.96 | "Aw -- take a walk!" |
| other | -1.97 | I'll learn him!" |
| other | -1.98 | "By jingo! |
| other | -2.11 | "You can't." |
| other | -2.13 | Course you would!" |
| other | -2.16 | "Y-o-u-u TOM!" |
| other | -2.17 | Oh, what a hat!" |
| other | -2.17 | "Well why don't you?" |
| other | -2.18 | Why don't you DO it? |
| other | -2.99 | "Nothing! |
| other | -3.20 | Ting-a-ling-ling! |
| other | -3.20 | Chow-ow-ow! |
| other | -3.20 | Ting-a-ling-ling! |
| other | -3.20 | SH'T! |

Table 4: Predicted PC1 value and corresponding text for some sentences of chapter 01.


http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, June. Association for Computational Linguistics.

Christopher Potts and Florian Schwarz. 2010. Affective 'this'. *Linguistic Issues in Language Technology*, 3(5):1–30.

Christopher Potts. 2011a. On the negativity of negation. In Nan Li and David Lutz, editors, *Proceedings of Semantics and Linguistic Theory 20*, pages 636–659. CLC Publications, Ithaca, NY.

Christopher Potts. 2011b. Sentiment Symposium Tutorial: Lexicons. Section 3.4 Experience Project reaction distributions. http://sentiment.christopherpotts.net/lexicons.

Project Gutenberg. 2012. Free eBooks by Project Gutenberg. http://www.gutenberg.org/wiki/Main_Page.

R Development Core Team, 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

J. Reilly and L. Seibert. 2003. Language and emotion. In Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors, *Handbook of Affective Sciences*, chapter 27, pages 535–559. Academic Press.

K. Sjölander. 2012. The snack sound toolkit. http://www.speech.kth.se/snack.

E. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen. 2011. Clustering expressive speech styles in audiobooks using glottal source paramters. In *Interspeech*, Florence, Italy.

L. Wang, Y. Zhao, M. Chu, Y. Chen, F. Soong, and Z. Cao. 2006. Exploring expressive speech space in an audiobook. In *Speech Prosody 2006*, Dresden, Germany.