CORRELATIONS BETWEEN POSITIVE OR NEGATIVE UTTERANCES AND BASIC ACOUSTIC FEATURES OF VOICE: A PRELIMINARY ANALYSIS

ŁUKASZ STOLARSKI
Jan Kochanowski University in Kielce, Poland
lukasz.stolarski@ujk.edu.pl

Abstract
The major aim of this paper is to establish possible correlations between continuous sentiment scores and four basic acoustic characteristics of voice. In order to achieve this objective, the text of "A Christmas Carol" by Charles Dickens was tokenized at the sentence level. Next, each of the resulting text units was assessed in terms of sentiment polarity and aligned with the corresponding fragment in an audiobook. The results indicate weak but statistically significant correlations between sentiment scores and three acoustic features: the mean F0, the standard deviation of F0 and the mean intensity. These findings may be useful in selecting optimal acoustic features for model training in multimodal sentiment analysis. Also, they are essential from a linguistic point of view and could be applied in studies on such language phenomena as irony.

Key words: sentiment analysis, acoustic features, feature selection

1. Aims

Estimating the sentiment value of a given linguistic expression is typically performed on the basis of its written representation. Additionally, in the approach called "multimodal sentiment analysis" (see Section 0) other "modalities" are used, such as visual and acoustic content. This article focuses on the acoustic aspect, which has not been properly investigated from the linguistic point of view. The lack of such research may result from the engineering approach followed by specialists designing sentiment analysis software. It is paramount for such computer programs to perform well, and theoretical investigations on the subject are of secondary importance. Additionally, in many cases programmers have limited linguistic knowledge and, on the other hand, linguists are unable to use software packages without graphic user interfaces. As a consequence, proper theoretical investigation of the subject is not undertaken by either of the two groups.

The present article seeks to fill this gap by exploring possible correlations between continuous sentiment scores and four basic acoustic features: fundamental voice frequency, the variability of the fundamental frequency, voice intensity and the variability of voice intensity. Even though this analysis should be treated as preliminary and more research is needed in which other acoustic

characteristics of voice would be analysed, the results could be useful for programmers dealing with sentiment analysis. Furthermore, the findings of this study shed light on the purely linguistic question of how semantic content may affect articulation.

2. Background

Sentiment analysis deals with "the polarity of an opinion item which either can be positive, neutral or negative" (Borth, Ji, Chen, Breuel and Chang, 2013: 223). Hutto and Gilbert narrow down the meaning of the term to "an active area of study in the field of natural language processing that analyses people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text" (2014: 217). Likewise, Pang and Lee describe it as "the computational treatment of opinion, sentiment, and subjectivity in text" (2008: 10). The topic is currently widely discussed in publications on natural language processing, and several reviews summarizing the literature on sentiment analysis are available (e.g. Liu, 2012; Liu and Zhang, 2012; Pang and Lee, 2008).

As outlined in Stolarski (2021), one may distinguish two main categories of tools for performing sentiment analysis. Firstly, a number of software packages make use of the "machine learning approach" (Ribeiro, Araújo, Gonçalves, Gonçalves and Benevenuto, 2016; Taboada, Brooke, Tofiloski, Voll and Stede, 2011). Such programs involve classifiers build on the basis of labelled training data. They are useful for specific types of texts because they perform well in the domain that they were trained on. Nevertheless, they are much less reliable in other domains (Aue and Gamon, 2005) and they do not cope well with negation or intensification (Taboada et al., 2011). An alternative approach involves "sentiment lexicons", in which individual words, or sometimes phrases, are assigned sentiment values. These values could be categorical, as in the NRC Emotion Lexicon (Mohammad and Turney, 2010, 2013), or continuous, as in those lexicons tested in Stolarski (2021), e.g. the NRC Hashtag Sentiment Lexicon (Kiritchenko, Zhu and Mohammad, 2014; Mohammad, Kiritchenko and Zhu, 2013; Zhu, Kiritchenko and Mohammad, 2014), the Sentiment Composition Lexicon for Opposing Polarity Phrases (Kiritchenko and Mohammad, 2016a, 2016b), the SenticNet project lexicon (Cambria, Poria, Bajpai and Schuller, 2016), the SentiStrength lexicon (Thelwall, 2017; Thelwall et al., 2013; Thelwall, Buckley and Paltoglou, 2012; Thelwall, Buckley, Paltoglou, Cai and Kappas, 2010; Thelwall and Buckley, 2013) or in the VADER lexicon (see Sections 0 and Appendix). Tools using this methodology are usually more appropriate for general-purpose sentiment evaluation.

Although some sentiment analysis software packages are potentially more accurate than others, several benchmark comparisons have shown that no clear method is preferable for all possible testing sets (Abbasi, Hassan and Dhar, 2014;

Diniz et al., 2016; Gonçalves, Araújo, Benevenuto and Cha, 2013; Ribeiro et al., 2016). As a consequence, the performance of a particular tool needs to be evaluated on the type of texts for which it is to be ultimately used.

The use of acoustic features is common in the so-called "multimodal sentiment analysis". This is a special type of sentiment analysis which, in addition to textual content, uses other "modalities". These include acoustic measurements (Aldeneh, Khorram, Dimitriadis and Provost, 2017; Govindaraj and Gopalakrishnan, 2016; Li, Dimitriadis and Stolcke, 2019: 2; Sheikh, Dumpala, Chakraborty and Kopparapu, 2018) and also visual content (Chen et al., 2017, 2017; Pereira, Pádua, Pereira, Benevenuto and Dalip, 2016; Pérez-Rosas, Mihalcea and Morency, 2013; Rosas, Mihalcea and Morency, 2013; Wöllmer et al., 2013; Zadeh, Chen, Poria, Cambria and Morency, 2017). The acoustic modality typically involves either several, mostly randomly chosen, acoustic cues, or sets of features assembled for specific tasks offered in software packages such as openSMILE (Eyben, Weninger, Gross and Schuller, 2013). These acoustic measurements, however, are treated as tools for obtaining more accurate results and are not really investigated themselves (for more details see the discussion in Section 0). This is true even for publications which seemingly focus on acoustic features. For instance, Mairesse, Polifroni and Di Fabbrizio (2012) investigate the association between the results of a sentiment classification procedure and a set of "prosodic" voice characteristics. The summary provided shows that, indeed, the use of such features improves classification accuracy, but no details are given on the contribution of individual acoustic dimensions. The same is true for the analysis involving a set of voice features presented in Peng (2017).

Individual acoustic characteristics of the human voice have been, however, extensively discussed in relation to emotions. General surveys of studies dealing with this subject are provided in Frick (1985), Kappas, Hess, and Scherer (1991) and Scherer (1986). It is important to stress that affective speech is indirectly related to the major topic of this paper because some emotion categories may be interpreted as positive (e.g. joy or happiness), and others may be interpreted as negative (e.g. sadness or anger). As a consequence, one may formulate initial assumptions about acoustic characteristics of positive and negative utterances on the basis of the way some emotion dichotomies are conveyed acoustically. In the remaining part of this section, reports which deal with acoustic features and emotion categories relevant to the present study will be discussed.

The fundamental frequency is the most frequently described acoustic feature in relation to affective speech. In fact, the body of literature dealing with this particular topic is too great to be covered in detail here. Thus, the way the dichotomy "happy vs. sad" tends to be conveyed acoustically, as it is summarized in Stolarski (2020), will be presented as an example. These two emotion categories provide a good "positive-negative" contrast. It has been observed that the mean fundamental frequency tends to be higher in expressions involving happiness (Coleman and Williams, 1979; Collier and Hubbard, 1998, 2001; Davitz, 1964; Skinner, 1935), and lower in sad utterances (Fonagy, 1978; Huron, 2008; Huron,

Yim and Chordia, 2010; Leinonen, Hiltunen, Linnankoski and Laakso, 1997; Razak, Abidin and Komiya, 2003; Sobin and Alpert, 1999; Wallbott and Scherer, 1986). What is more, the variability of the fundamental frequency tends to be higher in expressions involving happiness (Breitenstein, Van Lancker and Daum, 2001; Fairbanks and Pronovost, 1939; Kaiser, 1962; Ladd, Silverman, Tolkmitt, Bergmann and Scherer, 1985; Wu, Zheng, Xu and Bao, 2006), and lower in utterances involving sadness (Breitenstein et al., 2001; Ladd et al., 1985; Skinner, 1935; Wu et al., 2006). On the basis of these findings, one may assume that the mean fundamental frequency should be higher in sentences with a high sentiment score, and lower in sentences with a low sentiment score. Likewise, the standard deviation of fundamental frequency should also be associated with similar tendencies. It should be higher in positive utterances, and lower in negative utterances.

As far as the intensity of sound waves is concerned, it has been reported that emotion categories, such as joy or elation, are frequently associated with raised mean sound pressure level (Bezooijen, 1984; Davitz, 1964; Höffe, 1960; Huttar, 1968; Kaiser, 1962; Skinner, 1935) and sadness or dejection with lowered mean sound pressure level (Davitz, 1964; Eldred and Price, 1958; Hargreaves, Starkweather and Blacker, 1965; Huttar, 1968; Kaiser, 1962; Skinner, 1935; Zuberbier, 1957). According to these results, one may expect positive utterances to have higher intensity of sound waves than negative utterances. It must be pointed out, however, that there is an alternative interpretation of these findings proposed in Stolarski (2020). Namely, shifts in sound pressure level could be dependent more on the intensity of a given emotion than on its kind. This line of reasoning is substantiated by the fact that emotion categories such as rage or hot anger are also associated with increased voice intensity (see the summaries in Kappas et al. 1991 and Scherer 1986) even though they may be interpreted as the opposite of joy or elation. This apparent inconsistency could be explained by referring to the model of emotions proposed by Plutchik (1980, 1997, 2000, 2001a, 2001b), in which rage is the same emotion as sadness, but it is more intense. In fact, the model treats some emotion labels as being on the same dimension, but differing in terms of intensity. This interpretation makes the predictions of acoustic characteristics of positive and negative utterances less definitive.

3. Methods

In order to accomplish the aims outlined in Section 0, a representative number of text fragments need to be assessed in terms of their sentiment value. Moreover, these excerpts must be read by a native speaker of English. Next, the resulting recordings should be analysed in terms of selected acoustic features. Finally, possible correlations between the sentiment values and the corresponding acoustic

values are to be investigated. The sections below explain the way in which these tasks are handled in this project.

3.1. Materials

The texts used in the present analysis could have been selected from a variety of genres, potentially involves differing degrees of sentiment value. The decision to choose texts from a novel was taken for several reasons. Firstly, there is a large selection of e-books freely available for download on websites such as "gutenberg.org". Secondly, many corresponding audiobooks may also be found, so the additional task of recording someone's voice is simplified. Finally, prose is usually associated with certain amount of emotional content, which adds to the value of this choice.

The text chosen for the present analysis is the novella "A Christmas Carol" by Charles Dickens. It contains 28,527 word tokens, which is a large enough sample for the present purposes. The text is available at "gutenberg.org". Additionally, the corresponding audiobook may be obtained at "librivox.org", which is a non-profit library of free audiobooks recorded by volunteers. In fact, the site offers as many as 11 different versions of "A Christmas Carol". The one which has been chosen here is read by a male speaker of General American English. It was downloaded in 5 separate mp3 files, each one with a bit rate of 64 kbit/s.

3.2. Text tokenization

The text of "A Christmas Carol" had to be divided into semantically and prosodically independent units. Whole chapters and paragraphs would have been too long and such tokenization would also have resulted in a smaller number of examples for analysis. Phrases and single words, on the other hand, would have been too short. Moreover, in many cases, they are semantically dependent on the textual and situation context. This is especially true for function words. In addition, phrases and single words are not independent prosodic entities, unless they are pronounced in isolation. Consequently, the option chosen for this project applies tokenization at the sentence level. This level frequently correlates with semantic and prosodic boundaries, although longer sentences consisting of several clauses may exhibit articulatory characteristics similar to a set of shorter sentences.

The tokenization was performed in Python using the "nltk.tokenize" package. It was appropriate for narration, but additional processing was necessary for dialogues. The problem is presented in the example below:

> "What else can I be," returned the uncle, "when I live in such a world of fools as this?" ("A Christmas Carol", Stave 1)

Scrooge's utterance is interrupted by a narrator's comment. From the orthographic point of view, the whole expression is one sentence, but semantically and prosodically we are dealing with two separate sentences. Such cases are relatively frequent in fiction and needed to be addressed. A similar problem was identified in Stolarski (2018). It was solved by writing an additional script in Python that divided such examples into separate parts. The same program was used for the current study. The script applies punctuation conventions used with dialogues in English. In the example above, the resulting units are:

1. What else can I be,
2. returned the uncle,
3. when I live in such a world of fools as this?

3.3. Text classification

Tokenization performed in the way specified in Section 0 makes it possible to categorize individual texts according to the pragmatic categories of meaning proposed by Huddleston (1988): statement, directive, question and exclamatory statement. In Speech Act Theory (Austin, 1962), these categories refer to the notion of "illocutionary force" and are prototypically associated with grammatical forms in the way presented in Table 1. These associations are encountered in direct speech acts but can be altered in indirect speech acts. The automatic categorization performed on the tokenized text refers to these categories of meaning rather than grammatical forms for the following two reasons:

1. The script which performed the categorization makes use of punctuation conventions of the English language. These conventions reflect categories of meaning rather than grammatical forms. For example, the question mark indicates that a given expression is meant to be a question, even if the grammatical form used is not an interrogative. In spoken language, this information is usually conveyed through an appropriate intonation pattern. Likewise, the exclamation mark signals that a given expression is meant to be an exclamatory statement, regardless of the grammatical form.
2. Meaning tends to affect prosodic aspects of voice more than grammatical form does. For instance, a text ending with a question mark will be read with one of the intonation patterns typical for a question, even if the grammatical form is declarative, imperative or exclamative.

Any textual units ending with a question mark were classified as questions, and any textual units ending with an exclamation mark were categorized as exclamatory statements. Unfortunately, there was no reliable way to discriminate

between the other two categories of meaning. Consequently, the categories applied are statements/directives, questions and exclamatory statements.

As will be shown in Section 0, meaning classification is very useful for the purposes of the present study. The acoustic response variables specified in Section 0 are affected by this factor to a significant degree, so a separate analysis for each meaning category is preferable. More importantly, however, this classification makes it possible to reduce noise in sentiment scoring. For instance, questions are not assessable as true or false and, usually, cannot be reliably graded for sentiment polarity. Therefore, excluding questions may result in more accurate sentiment values.

Table 1: Prototypical associations between grammatical forms and categories of meaning in direct speech acts

Grammatical form	Category of meaning (illocutionary force)
Declarative	Statement
Imperative	Directive
Interrogative	Question
Exclamative	Exclamatory Statement

Another group of examples which are problematic for the purposes of the present study are texts which involve negation. This concerns both "verb negation" (structures such as *do not, don't, hasn't*, etc.) and "non-verb negation" (words such as *never*, *nobody*, etc.), to use the terminology employed in Huddleston (1988). In sentiment analysis projects the issue was initially addressed by reversing the polarity of a lexical item (Choi and Cardie, 2008; Kennedy and Inkpen, 2005), but this approach has been shown to be fundamentally flawed (Kennedy and Inkpen, 2006; Kiritchenko et al., 2014; Taboada et al., 2011). Although later solutions, such as shifting the polarity by a fixed amount (Taboada et al., 2011), yield more promising results, negation is still a challenge in sentiment analysis and only some tools are designed to deal with it. The program used in the present analysis (see Section 0) is capable of processing texts with "verb negation". It must be stressed, however, that the precision of sentiment scoring may be lower for such examples and, additionally, the program does not cope with "non-verb negation". All this points to the fact that examples involving negation should be treated with caution. As a result, another python script was written which classified all text fragments according to whether or not they included negation.

3.4. Text sentiment analysis

As indicated in Section 0, there is a wide choice of software packages designed for sentiment analysis. Even though their performance tends to differ depending on a particular testing dataset and in validation tests no clear winner may be established, some are more suitable for the purposes of this study than others. The most appropriate sentiment evaluation program should be a general-purpose tool designed to deal with different types of text. This excludes software packages within the "machine learning approach", which are usually used for conducting sentiment analysis on texts in a particular domain. Instead, a lexicon-based tool should be considered. The software package which has been selected for the present project is "Valence Aware Dictionary and sEntiment Reasoner", or VADER (Hutto and Gilbert, 2014). It is especially attuned to social media contexts but may also be applied to other domains. Additionally, the program offers numerical scoring. This is particularly useful since it enables assessing the relative strength of sentiment polarity instead of just indicating that a given text is positive or negative.

The software is available as a Python library. Consequently, another script in Python was written in which each text unit obtained after the tokenization described Section 0 was assessed using VADER on a scale of -1 (very negative) to + 1 (very positive).

3.5. Text-audio alignment

"Forced alignment" is the process of automatic synchronization of audio and text. It may be performed using various software packages, such as Julius (Lee and Kawahara, 2009), EasyAlign (Goldman, 2011) or Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner and Sonderegger, 2017). The particular tool chosen for the present project is called "Aeneas". It is a Python/C library which is designed to perform forced alignment at the sentence level. It supports multiple output options, including Praat's TextGrid format. Consequently, the tool is optimal for the purposes of the present study.

Before performing forced alignment on the materials described in Section 0, some manual pre-processing was necessary. The audiobook of "A Christmas Carol" was downloaded in the form of five separate mp3 files, each referring to a separate chapter, or "stave", of the novella. In each recording, the reader made remarks before and after reading the proper text. These additional comments were removed using Audacity (Audacity Team, 2014). Likewise, the text downloaded from "gutenberg.org" as a "txt" file contained parts which did not belong to the novella itself (e.g. general information about Project Gutenberg and a copyright notice). These fragments were also removed. Finally, forced alignment was performed on the text tokenized in the way described in Section 0.

3.6. Acoustic analysis

The acoustic features of speech which have been chosen as dependent variables are the arithmetic mean fundamental frequency (measured in Hz using deafault settings in Praat), the variability of the fundamental frequency (measured as the standard deviation of the fundamental frequency), the arithmentic mean voice intensity (measured in dB) and the variability of voice intensity (measured as the standard deviation of voice intensity). They will be abbreviated as F0, SD of F0, INT and SD of INT, respectively. There are two reasons why these features were chosen. Firstly, they are among the basic acoustic characteristics of the human voice, so their choice is appropriate for a preliminary study on acoustic correlates of sentiment. Secondly, as described in Section 0, in studies on affect vocalization, such basic features are of special interest as they are assumed to be potentially more influenced by emotions than many other features. Since studies on the vocal expression of affect are the major source of predictions for the current research, the choice of F0, SD of F0, INT and SD of INT seemed optimal.

In order to measure the values of these acoustic features for each fragment of the audiobook aligned with each tokenized text unit (see Section 0), a Praat (Boersma and Weenink, 2014) script was written. The script took the audio files downloaded from "librivox.org" and the TextGrid files generated using Aeneas (see Section 0) as input. The results obtained were saved as a "tsv" file. Finally, with the use of a new Python script, acoustic data were matched with corresponding numeric sentiment values that had been generated using VADER (see Section 0).

3.7. Statistical analysis

The database created for this project contains 2,271 entries. An example of the first 15 is presented in Table **Błąd! Nie można odnaleźć źródła odwołania.**. Each entry contains a text unit which was extracted from "A Christmas Carol" in the process of tokenization described in Section 0. Next, the illocutionary force is specified and the category "positive" or "negation" is given. The rest of the data are numeric. This includes sentiment scores obtained using VADER (see Section 0) and the values for each of the acoustic features discussed in Section 0.

The sentiment scores are treated as the only independent, or predictor, numeric variable. The acoustic measurements, on the other hand, are dependent variables, whose values are potentially affected by the sentiment scores.

Table 2: First 15 entries in the database created for the current project

Text Unit / Audio Unit	Text Categorization			Acoustic Measurements			
	Illocution	Positive vs. Negation	VADER score	F0 (Hz)	SD of F0 (Hz)	INT (dB)	SD of INT (dB)
Marley was dead, to begin with.	statement/directive	positive	-0.6486	113.06	26.86	45.26	25.71
There is no doubt whatever about that.	statement/directive	negation	-0.5719	135.49	39.45	57.4	18.63
The register of his burial was signed by the clergyman, the clerk, the undertaker, and the chief mourner.	statement/directive	positive	-0.3818	131.4	27.63	54.02	20.39
Scrooge signed it.	statement/directive	positive	0	131.22	37.02	46.87	22.74
And Scrooge's name was good upon 'Change for anything he chose to put his hand to.	statement/directive	positive	0.7269	126.26	26.38	55.43	21.28
Old Marley was as dead as a door-nail.	statement/directive	positive	-0.6486	147.34	72.94	53.6	21.49
Mind!	exclamatory statement	positive	0	126.32	15.74	42.53	27.97
I don't mean to say that I know, of my own knowledge, what there is particularly dead about a door-nail.	statement/directive	negation	-0.6801	122.34	21.53	57.05	20.17
I might have been inclined, myself, to regard a coffin-nail as the deadest piece of ironmongery in the trade.	statement/directive	positive	0	119.4	24.58	56.65	19.35
But the wisdom of our ancestors is in the simile; and my unhallowed hands shall not disturb it, or the Country's done for.	statement/directive	negation	0.6866	137.1	32.99	53.44	22.6
You will, therefore, permit me to repeat, emphatically, that Marley was as dead as a door-nail.	statement/directive	positive	-0.6486	138.47	35.42	55.27	19.4
Scrooge knew he was dead?	question	positive	-0.6486	138.08	37.17	57.48	18.84
Of course he did.	statement/directive	positive	0	142.82	93.56	47.94	18.62
How could it be otherwise?	question	positive	0	108.98	26.15	51.11	21.01

The major statistical aim of this project is to investigate correlations between the sentiment scores and the values for each of the acoustic features. In order to choose optimal methods, however, the normality of the samples under analysis needed to be considered. Figure 1 provides histograms for all the numeric datasets used in the present study. It is clear that they do not strictly follow a normal distribution. Most of the sentiment scores obtained using VADER are slightly below 0, and the distribution of more negative and all positive values deviates significantly from the Gaussian ideal, indicating a leptokurtic distribution (the value of kurtosis for this dataset is 3.0406). Additionally, the results of both Shapiro-Wilk ($W = 0.93127$, p-value < 0.0001) and Anderson-Darling ($A = 89.45$, p-value < 0.0001) normality tests indicate that the sample does not have a normal distribution. The histograms for acoustic measurements resemble a bell-shaped curve. Nevertheless, their distributions are visibly skewed, either to the right (in the case of F0 and SD of F0), or to the left (as in the case of INT and SD of INT), indicating outliers. This is also confirmed by the skeweness test, which yielded values of 1.2488, 1.2735, -0.8425 and -0.5214, respectively. Finally, normality tests performed on these datasets indicate p-values below 0.0001.

Such results suggest that correlations between VADER scoring and the other numeric datasets should be evaluated using the non-parametric Spearman's rank correlation coefficient. Consequently, this statistic will be given priority in

Section 0, although the Pearson correlation coefficient will also be considered to a limited extent.

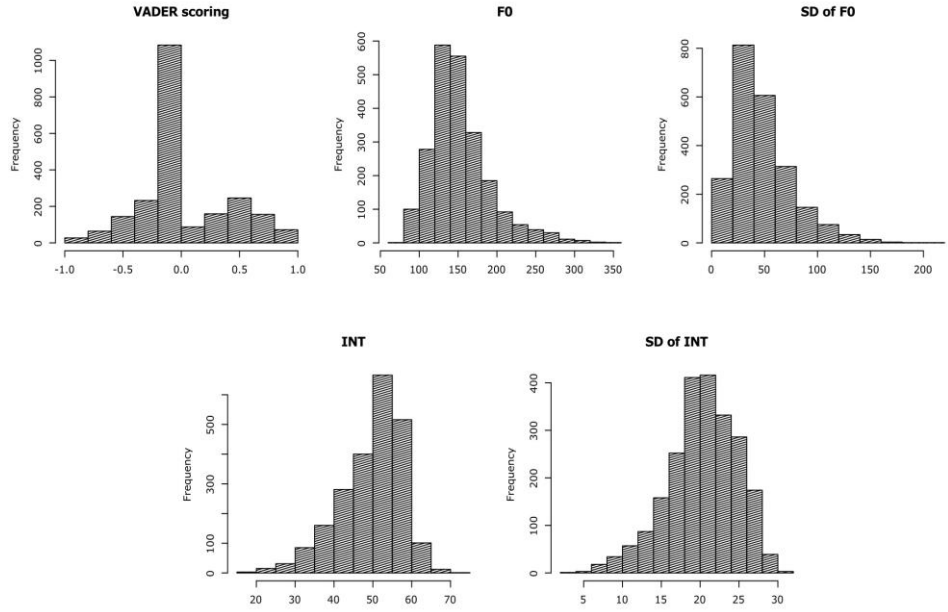

Figure 1: Histograms for all numeric datasets used in the analysis

All the statistical tests in this study were performed with the use of *R 3.4.4* (R Development Core Team, 2018).

4. Results

The results are presented in four separate sections. The first one provides an analysis of the whole sample. The following three parts focus on contexts in which sentiment scoring obtained from VADER is potentially more accurate than in the entire dataset. In Section 0 examples with negation are excluded. The analysis in Section 0 reveals that excluding questions and exclamative statements in addition to negation results in stronger correlation coefficients. Finally, in Section 0 only the optimal cases are selected. In this analysis, texts with low absolute sentiment scores are also excluded.

4.1. Results based on the whole sample

Figure 2 shows scatterplots of VADER sentiment scores juxtaposed against the dependent acoustic variables specified in Section 0. It also provides Pearson's product-moment correlation coefficients and, more importantly, Spearman's rank correlation coefficients. It is clear that the correlations are very weak. For instance, Pearson's correlation coefficient for VADER scores juxtaposed against measurements of F0 is 0.073. The corresponding Spearman's rho is only slightly higher (0.095). These measurements are confirmed by the linear regression line represented by the dashed line in the middle of the scatterplot in the upper left-hand corner of Figure 2. The line raises only slightly. Still, because of the large sample used, these results are statistically significant (p-values are below 0.001 in both cases). Similar observations can be made regarding SD of F0 and INT. Again, the correlation coefficients obtained are very small and the regression lines barely deviate from the horizontal position, but the corresponding p-values indicate that the correlations are statistically relevant. The results for SD of INT, on the other hand, do not suggest any possible correlation. Both r and rho are close to 0 and the p-values obtained in both tests are clearly above the alpha level of 0.05 (see the scatterplot in the lower right-hand corner of Figure 2).

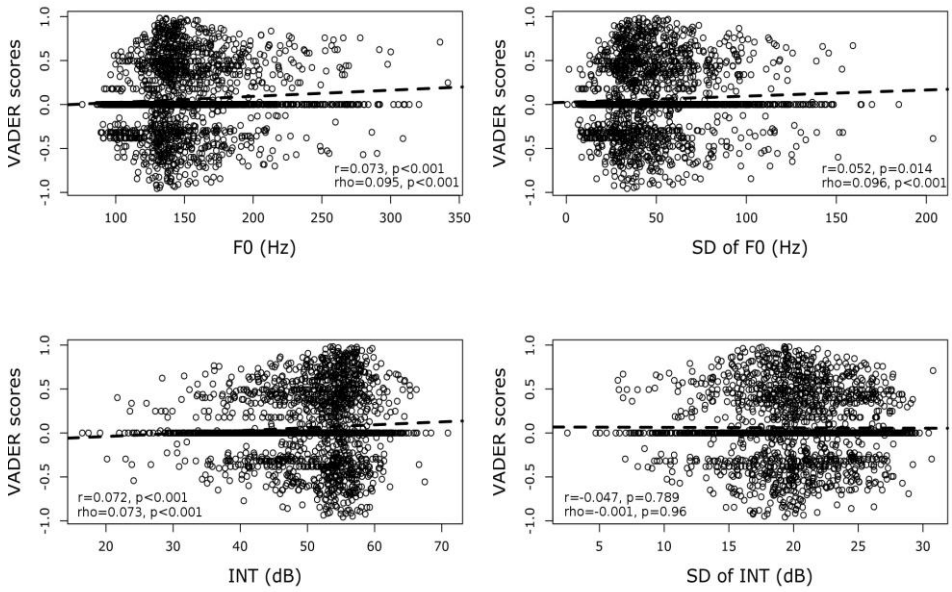

Figure 2: Scatterplots of VADER sentiment scores juxtaposed against the dependent acoustic variables under analysis (in each case, n=2271)

It is worth mentioning that attempts to normalize the acoustic measurements did not improve the results. With the use of a script written in R, the mean value of a given acoustic variable was deducted from the corresponding acoustic measurements. For example, from each F0 measurement the mean F0 obtained on the basis of all F0 values in the database was deducted. In this way it was possible to analyse values representing relative changes in F0 rather than absolute measurements in Hz. Such transformations, however, did not affect the results in any way. Likewise, additional modifications by which these relative differences were expressed in percentage terms also did not change the outcome of correlation tests. The reason why speaker normalization did not change the results in this case was the fact that only one speaker is being investigated. Therefore, it did not matter which units of measurement were used. If more readers had been analysed, however, such normalization procedures would have been necessary to obtain accurate results.

4.2. Results excluding negation

According to the results of the validation tests performed on VADER (see Appendix), the precision of sentiment scoring is slightly lower in texts involving negation. Consequently, excluding such examples may decrease the amount of noise in the data and, eventually, allow for better detection of possible trends. This assumption is confirmed in Table **Błąd! Nie można odnaleźć źródła odwołania.** which presents correlation coefficients for all acoustic values under analysis juxtaposed against VADER scores obtained only in texts which did not involve negation. The Pearson's correlation coefficients are higher than those obtained in Section 0 by about 0.02, except for the result concerning SD of INT, in which case, the lack of correlation is indicated by an even lower value (r=-0.007). Moreover, the Spearman's correlation coefficients increase to an even greater extent. For instance, the value obtained for F0 is 0.122, which is 0.027 more than in the result yielded in Section 0. Similar relative changes are observable for SD of F0 and INT, but, again, the value of rho for SD of INT is close to 0.

Table 3: Correlation coefficients for the data without examples involving negation, n=1904

	Pearson's correlation coefficient	P-value	Spearman's correlation coefficient	P-value
F0 – VADER	0.094	< 0.001	0.122	< 0.001
SD of F0 – VADER	0.065	0.004	0.115	< 0.001
INT – VADER	0.098	< 0.001	0.103	< 0.001
SD of INT – VADER	-0.007	0.748	0.001	0.979

4.3. The results across pragmatic categories of meaning

As described in Section 0, text units were classified into three categories of meaning: statements/directives, questions and exclamative statements. According to the results of the validation tests performed on VADER (see Appendix), this division may be used to obtain more accurate results. It was observed that the sentiment scoring yielded in VADER is more precise when questions are excluded from the analysis. This was predictable because questions are generally problematic in sentiment analysis. As discussed in Section 0, they cannot be assessed as true or false. Moreover, analysing statements/directives, questions and exclamative statements separately makes sense since this distinction directly affects acoustic response variables. An example is shown in Figure 3 in which the values of F0 are separated by the three levels of pragmatic meaning. It is visible that the category of statements/directives tends to be associated with relatively lower mean F0 than the other two categories. One-way ANOVA performed on these data yields a p-value smaller than 0.001 and Tukey's multiple comparisons test indicates that the difference between statements/directives and exclamative statements, as well as between statements/directives and questions, is statistically significant. A precise analysis of the effects of pragmatic meaning categories on acoustic measurements is beyond the scope of this study, but such results suggest that this factor should be controlled for.

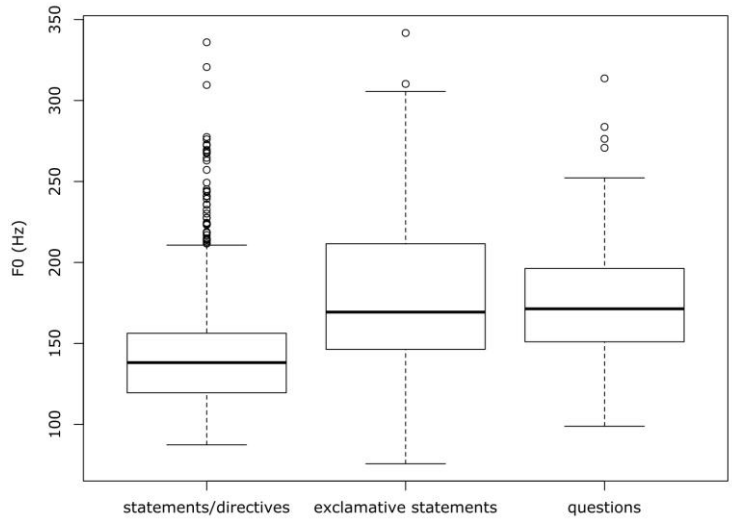

Figure 3: The reader's F0 across statements/directives, exclamative statements and questions

Table **Błąd! Nie można odnaleźć źródła odwołania.** summarises the results of correlation tests performed on statements/directives only (negation was also excluded from these data). The values obtained are noticeably larger than the ones reported in Table **Błąd! Nie można odnaleźć źródła odwołania.**. The average improvement for Pearson's correlation coefficients is about 0.035, except for SD of INT, which, again, is not correlated with sentiment scores. The increase in the Spearman's correlation coefficients is even larger. On average, it is about 0.048, resulting in values around 0.16. Once again, this does not concern SD of INT, which is not affected by the sentiment factor.

Table 4: Correlation coefficients for statements/directives only (examples with negation also excluded), n=1327

	Pearson's correlation coefficient	P-value	Spearman's correlation coefficient	P-value
F0 – VADER	0.133	< 0.001	0.168	< 0.001
SD of F0 – VADER	0.085	0.002	0.157	< 0.001
INT – VADER	0.143	< 0.001	0.158	< 0.001
SD of INT – VADER	-0.018	0.515	-0.015	0.587

Tables **Błąd! Nie można odnaleźć źródła odwołania.** and **Błąd! Nie można odnaleźć źródła odwołania.** present the results of correlation tests performed on exclamatory statements and questions, respectively. In all these cases, the correlation coefficients are close to 0 and the corresponding p-values are clearly above the alpha level of 0.05. Such findings are not very surprising for questions, but they are for exclamative statements. One possible explanation of this result is that the exclamative statements found in "A Christmas Carol" are very short in comparison with the length of statements/directives. While the mean length of the latter is 12.2 words, the average length of exclamative statements is only 6.8 words. For this reason, the sentiment scores obtained in VADER may be less precise for exclamative statements than for other groups. Another reason could be due to the relatively smaller sample. Exclamative statements are represented by 401 examples, while statements/directives comprise 1327 text units. Nevertheless, regardless of the actual reasons, neither questions nor exclamative statements are correlated with sentiment scores and in order to further investigate the way in which sentiment is conveyed acoustically they are excluded from the sample.

According to the results of the validation tests performed on VADER (see Appendix), sentiment scoring is more accurate in the cases in which texts which are clearly positive or negative. In fact, an analysis performed on examples whose sentiment scoring was below -1 standard deviation and above +1 standard deviation showed that the precision of sentiment values was close to the human level. This strategy may be used to further remove noise from the data and improve the detection of existing tendencies.

Table 5: Correlation coefficients for exclamatory statements only (examples with negation are also excluded), n=401

	Pearson's correlation coefficient	P-value	Spearman's correlation coefficient	P-value
F0 – VADER	-0.031	0.538	-0.048	0.342
SD of F0 – VADER	-0.013	0.787	-0.022	0.664
INT – VADER	0.048	0.335	0.023	0.645
SD of INT – VADER	-0.016	0.753	-0.021	0.671

Table 6: Correlation coefficients for questions only (examples with negation are also excluded), n=176

	Pearson's correlation coefficient	P-value	Spearman's correlation coefficient	P-value
F0 – VADER	-0.025	0.737	-0.033	0.667
SD of F0 – VADER	0.001	0.99	-0.011	0.88
INT – VADER	-0.057	0.453	-0.07	0.354
SD of INT – VADER	0.072	0.342	0.066	0.383

4.4. Results excluding low absolute sentiment scores

Figure 4 depicts scatterplots similar to the ones presented in Section 0, but this time only statements/directives not involving negation and with sentiment scores below -1 sd and above + 1 sd were included. The resulting sample is much smaller (n=463), but the linear regression lines represented as dashed lines in the middle of the first three scatterplots indicate stronger correlations than previously. This is confirmed by higher correlation coefficients, also provided in Figure 4. For instance, the Pearson's correlation coefficient for F0 juxtaposed against VADER scores increased from 0.073 in Section 0 to 0.223. The value is three times higher. Likewise, the difference between the Spearman's correlation coefficients is significant (0.194 - 0.095 = 0.099). This indicates that the correlation is much stronger in the sample analysed in this part of the paper than in the entire dataset investigated in Section 0. Similar conclusions may be drawn for SD of F0. The increase of both Pearson's correlation coefficient and Spearman's correlation coefficient is above 0.07, resulting in values twice as large. Moreover, the corresponding data obtained for INT has also increased. The differences between the respective correlation coefficients reported in Figure 3 and Figure 4 exceeds 0.13. The resulting values are above 0.2.

All these improvements may, at least partially, result from a possible tendency for sentiment polarity to be expressed more consistently in the type of texts in the

sample analysed in this part of the paper. Still, the validation test summarised in Appendix suggests that the more precise sentiment scoring obtained in these examples is also a crucial factor and the higher correlation coefficients indicate the trends more accurately.

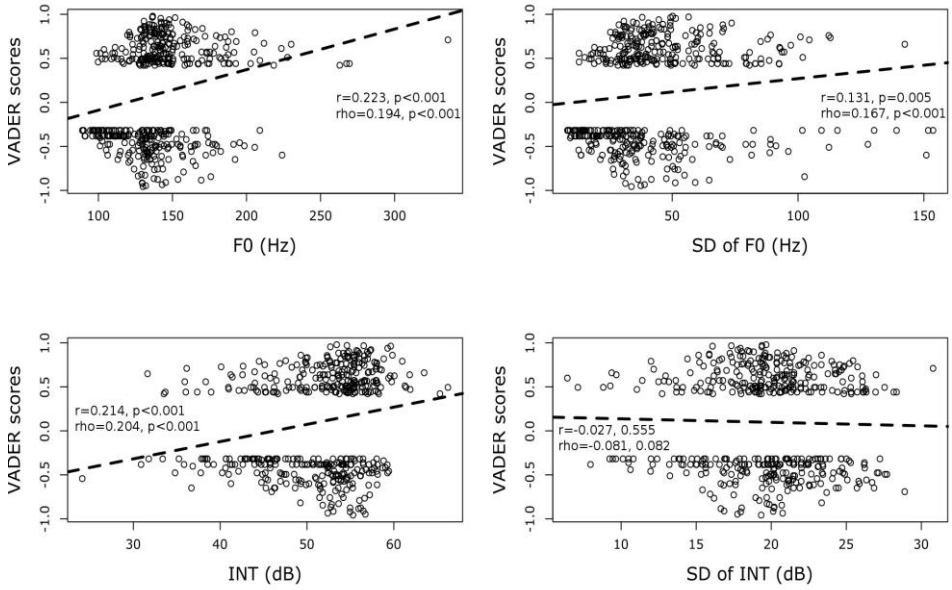

Figure 4: Scatterplots of VADER sentiment scores juxtaposed against the dependent acoustic variables under analysis. Only statements/directives are included. Additionally, texts with negation and sentiment scoring above -1 sd and below + 1 sd are excluded (in each case, n=463)

5. Discussion and conclusions

The initial results obtained on the basis of all the materials used in this study indicated statistically significant but very weak correlations between sentiment scores and F0, SD of F0 and INT (see Section 0). In subsequent sections examples which could potentially be problematic for VADER were gradually removed and the noise in the data was decreased. This resulted in incrementally higher correlation coefficients. Still, the general observation that three of the acoustic features under investigation correlated with sentiment scores, and SD of INT did not, was substantiated. In the final version of the analysis presented in Section 0 it was confirmed that F0 and INT could be helpful in acoustic detection of sentiment. The corresponding correlation coefficients were clearly above 0.2. SD of F0 could also be useful, although its relevance is relatively smaller. Spearman's correlation coefficient established for optimal example texts was 0.167. Finally, SD of INT was found not to be associated in any way with sentiment. The

correlation coefficients obtained were close to 0 and the corresponding p-values above the alpha level of 0.05.

These findings could be helpful in selecting acoustic features when training models for sentiment analysis. More importantly, the general approach of a systematic examination of selected features as presented in this study could be very useful for anyone dealing with acoustic features in the domain of natural language processing. Nevertheless, in order to substantiate these assumptions, further discussion is required.

As mentioned in Section 0, a standard strategy implemented in multimodal sentiment analysis, as well as in many other areas of computer technology dealing with sounds, is either to use a few, randomly chosen acoustic features or a set prepared for a specific purpose in packages such as openSMILE (Eyben et al., 2013). The advantages of the latter include a relatively larger number of acoustic features and reliance on the choice made by a group of programmers specializing in the field. However, there are also problems with such "custom sets". Firstly, sometimes several sets are available for performing one type of analysis. For example, openSMILE offers 2 different feature collections for speech recognition and as many as 4 sets for emotion recognition. Consequently, the user may feel confused as to which set is better for his/her project. Secondly, the predefined sets of acoustic features only cover a small fraction of potential speech analysis tasks. For instance, openSMILE does not provide any special collection of features for sentiment analysis. While sets prepared for emotion recognition may be used instead, they were not assembled with this particular task in mind. A possible solution is to use all available features. This is referred to as a "brute-force method" (Schuller et al., 2007). Indeed, some feature sets offered in openSMILE contain over 6,000 features and the use of collections as large as 50,000 is also reported (Schuller, Steidl and Batliner, 2009). This approach, however, generates new problems. For obvious reasons, numerous features included in such sets are not relevant for a particular task. This considerably increases the processing cost and potentially causes issues resulting from data sparsity in multi-dimensional datasets. The standard solution used to deal with these issues involves various automatic feature selection algorithms which eliminate features with lower information gain ratios. Still, it is not clear which of these algorithms to choose for a given project. Moreover, the classification accuracy of a model after such an automatic selection procedure may still be lower in comparison to the accuracy of a model trained on a smaller, but more appropriate set of features. An example may be found in Mairesse et al. (2012), where cross-validation tests indicate better results for sentiment classification based only on "F0 features" in comparison to all 988 features used in the analysis. A "knowledge-based approach" is, therefore, an important aspect, even if automatic feature selection procedures are used. Consequently, the description of the effects of individual features in a given type of task, such as acoustic sentiment analysis, is very helpful. It is better to exclude features which are inappropriate for a given project before training a model.

Another problem which should be addressed is that, regardless of the size of custom feature sets, some acoustic characteristics which should be included may be missing. This is evident from the fact that new features are constantly being added to the functionality of the software packages such as openSMILE, and feature sets have to be updated. Moreover, lack of understanding of individual acoustic characteristics may also affect the precision of measurements by not modifying additional parameters correctly. This happens, for instance, when obtaining values of F0. According to documentation on measuring pitch contours in Praat, the settings should involve different pitch floor and pitch ceiling values for men and women (75Hz-300Hz and 100-500Hz, respectively). Such adjustments are seldom applied when using large custom feature sets.

The results obtained in this study are also interesting from a purely linguistic point of view. They show that semantic content may affect articulation and, as a consequence, acoustic characteristics of speech. When saying something positive, speakers tend to raise their pitch (higher F0), apply more "prosodic explicitness" (SD of F0), as variability of F0 is called in Traunmüller and Eriksson (1995), and speak louder (INT). This resembles the phenomenon established for the aspects of pragmatic meaning described in Section 0. The categories statement, directive, question and exclamatory statement are associated with specific grammatical representations (declarative, imperative, interrogative and exclamative, respectively) and particular intonation patterns. There is certain amount of "redundancy", because the listener is provided with two independent but consistent cues, the grammatical form and the accompanying intonation pattern. The default relationships may be changed, however, and a given category of meaning may be conveyed by a different grammatical form. For instance, a directive may be expressed via an interrogative construction. This will result in additional "contextual effects", to use the terminology of Relevance Theory's conceptual framework (Sperber and Wilson, 1986). These effects would include, for example, the additional implicature that the speaker respects the listener and is trying to be polite. This line of reasoning could also be applied to the results of the present paper. Positive and negative utterances do not have separate grammatical forms but are explicitly expressed by lexical items with positive or negative semantic content. The articulatory dimension (and resulting acoustic characteristics) are superimposed on the lexical content. Again, there is some "redundancy", and the two levels, the lexical and the phonetic, participate in conveying the intended meaning to the listener. Nevertheless, the two dimensions may be matched differently, resulting in new contextual effects. The listener assumes that the speaker has expressed himself/herself in the most effective way (within the speaker's abilities) and by, for example, saying something positive in a voice which is typical for negative utterances, something additional is meant. In this case, the listener could infer that the speaker is using irony.

In future research, analyses similar to the one performed in this study could be conducted on many other acoustic features of voice. This would enhance the

understanding of the way in which sentiment polarity affects articulation and potentially facilitate the creation of multimodal sentiment analysis models.

References

Abbasi, Ahmed, Hassan, Ammar and Dhar, Milan.2014. Benchmarking Twitter Sentiment Analysis Tools. In *LREC*, Vol. 14, 26–31.

Aldeneh, Zakaria, Khorram, Soheil, Dimitriadis, Dimitrios and Provost, Emily Mower. 2017. Pooling acoustic and lexical features for the prediction of valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 68–72. ACM. https://doi.org/10.1145/3136755.3136760

Audacity Team. 2014. Audacity(R): Free audio editor and recorder (version 2.0.5) [computer software].

Aue, Anthony and Gamon, Michael. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, Vol. 1, 2–10. Citeseer.

Austin, John Langshaw. 1962. *How to do things with words*. Oxford: Clarendon Press.

Bezooijen, Renée. 1984. *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht, Netherlands: Foris Publications. https://doi.org/10.1515/9783110850390

Boersma, Paul and Weenink, David. 2014. Praat, a system for doing phonetics by computer (version 5.4.01) [computer software]. Amsterdam: University of Amsterdam.

Borth, Damian, Ji, Rongrong, Chen, Tao, Breuel, Thomas and Chang, Shih-Fu. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, 223–232. ACM. https://doi.org/10.1145/2502081.2502282

Breitenstein, Caterina, Van Lancker, Diana and Daum, Irene. 2001. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, *15*(1), 57–79. https://doi.org/10.1080/02699930126095

Cambria, Erik, Poria, Soujanya, Bajpai, Rajiv and Schuller, Björn W. 2016. SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. In *COLING*, 2666–2677.

Chen, Minghai, Wang, Sen, Liang, Paul Pu, Baltrušaitis, Tadas, Zadeh, Amir and Morency, Louis-Philippe. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 63–171. ACM. https://doi.org/10.1145/3136755.3136801

Choi, Yejin and Cardie, Claire. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 793–801). Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613816

Coleman, Robert F. and Williams, Robert. 1979. Identification of emotional states using perceptual and acoustic analyses. In *Transcript of the 8th Symposium: Care of the Professional Voice, Part I. The Voice Foundation, New York*.

Collier, William G. and Hubbard, Timothy L. 1998. Judgments of happiness, brightness, speed, and tempo change of auditory stimuli varying in pitch and tempo. *Psychomusicology*, *17*(1/2), 36–55. https://doi.org/10.1037/h0094060

Collier, William G. and Hubbard, Timothy L. 2001. Musical scales and evaluations of happiness and awkwardness: Effects of pitch, direction, and scale mode. *American Journal of Psychology*, *114*(3), 355–375. https://doi.org/10.2307/1423686

Davitz, Joel R. 1964. Auditory correlates of vocal expressions of emotional meaning. *The Communication of Emotional Meaning*, 101–112.

Diniz, Joao P., Bastos, Lucas, Soares, Elias, Ferreira, Miller, Ribeiro, Filipe and Benevenuto, Fabrıcio. 2016. ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis.

Eldred, Stanley H. and Price, Douglas B. 1958. A linguistic evaluation of feeling states in psychotherapy. *Psychiatry*, *21*(2), 115–121. https://doi.org/10.1080/00332747.1958.11023120

Eyben, Florian, Weninger, Felix, Gross, Florian and Schuller, Björn. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, 835–838. ACM. https://doi.org/10.1145/2502081.2502224

Fairbanks, Grant and Pronovost, Wilbert. 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, *6*(1), 87–104. https://doi.org/10.1080/03637753909374863

Fonagy, Ivan.1978. A new method of investigating the perception of prosodic features. *Language and Speech*, *21*(1), 34–49. https://doi.org/10.1177/002383097802100102

Frick, Robert W. 1985. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, *97*(3), 412–429. https://doi.org/10.1037/0033-2909.97.3.412

Goldman, Jean-Philippe. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. In *Proceedings of Interspeech*, 3233–3236. https://doi.org/10.21437/Interspeech.2011-815

Gonçalves, Pollyanna, Araújo, Matheus, Benevenuto, Fabrício and Cha, Meeyoung. 2013. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27–38). ACM. https://doi.org/10.1145/2512938.2512951

Govindaraj, Sureshkumar and Gopalakrishnan, Kumaravelan. 2016. Intensified sentiment analysis of customer product reviews using acoustic and textual features. *ETRI Journal*, *38*(3), 494–501. https://doi.org/10.4218/etrij.16.0115.0684

Hargreaves, William A., Starkweather, John A. and Blacker, K. H. 1965. Voice quality in depression. *Journal of Abnormal Psychology*, *70*(3), 218–220. https://doi.org/10.1037/h0022151

Höffe, Wilhelm L. 1960. Über Beziehungen von Sprachmelodie und Lautstärke. *Phonetica*, *5*(3–4), 129–159. https://doi.org/10.1159/000258054

Huddleston, Rodney. 1988. *English grammar: An outline*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166003

Huron, David. 2008. A comparison of average pitch height and interval size in major-and minor-key themes: Evidence consistent with affect-related pitch prosody. *Empirical Musicology Review*, *3*(2), 59–63. https://doi.org/10.18061/1811/31940

Huron, David, Yim, Gary, and Chordia, Parag. 2010. The effect of pitch exposure on sadness judgments: An association between sadness and lower than normal pitch. In *Proceedings of the 11th International Conference on Music Perception and Cognition*, 63–66.

Huttar, George L. 1968. Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech, Language, and Hearing Research*, *11*(3), 481–487. https://doi.org/10.1044/jshr.1103.481

Hutto, C. J. and Gilbert, Eric. 2014. Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, 216–255. Ann Arbor, MI. https://doi.org/10.1609/icwsm.v8i1.14550

Kaiser, L. 1962. Communication of affects by single vowels. *Synthese*, *14*(4), 300–319. https://doi.org/10.1007/BF00869311

Kappas, Arvid, Hess, Ursula and Scherer, Klaus R. 1991. Voice and emotion. In R. S. Feldman and B. Rim (eds.), *Fundamentals of nonverbal behavior* (pp. 200–238). Paris, France: Editions de la Maison des Sciences de l'Homme.

Kennedy, Alistair and Inkpen, Diana. 2005. Sentiment classification of movie and product reviews using contextual valence shifters. In *Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*. Ottawa, Ontario, Canada.

Kennedy, Alistair and Inkpen, Diana. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, *22*(2), 110–125. https://doi.org/10.1111/j.1467-8640.2006.00277.x

Kiritchenko, Svetlana and Mohammad, Saif M. 2016a. Happy Accident: A Sentiment Composition Lexicon for Opposing Polarity Phrases. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC)*. Portoro, Slovenia. https://doi.org/10.18653/v1/N16-1128

Kiritchenko, Svetlana and Mohammad, Saif M. 2016b. Sentiment composition of words with opposing polarities. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 1102–1108. San Diego, California. https://doi.org/10.18653/v1/N16-1128

Kiritchenko, Svetlana, Zhu, Xiaodan and Mohammad, Saif M. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762. https://doi.org/10.1613/jair.4272

Ladd, D. Robert, Silverman, Kim E.A., Tolkmitt, Frank, Bergmann, Günther and Scherer, Klaus R. 1985. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America*, *78*(2), 435–444. https://doi.org/10.1121/1.392466

Lee, Akinobu and Kawahara, Tatsuya. 2009. Recent development of open-source speech recognition engine julius. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, 131–137. Asia-Pacific Signal and Information Processing Association, 2009 Annual ….

Leinonen, Lea, Hiltunen, Tapio, Linnankoski, Ilkka and Laakso, Maija-Liisa. 1997. Expression of emotional–motivational connotations with a one-word utterance. *The Journal of the Acoustical Society of America*, *102*(3), 1853–1863. https://doi.org/10.1121/1.420109

Li, Bryan, Dimitriadis, Dimitrios and Stolcke, Andreas. 2019. Acoustic and Lexical Sentiment Analysis for Customer Service Calls. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5876–5880. IEEE. https://doi.org/10.1109/ICASSP.2019.8683679

Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.1007/978-3-031-02145-9

Liu, Bing and Zhang, Lei. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, 415–463. Springer. https://doi.org/10.1007/978-1-4614-3223-4_13

Mairesse, François, Polifroni, Joseph and Di Fabbrizio, Giuseppe. 2012. Can prosody inform sentiment analysis? experiments on short spoken reviews. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012* 5093–5096. IEEE. https://doi.org/10.1109/ICASSP.2012.6289066

McAuliffe, Michael, Socolof, Michaela, Mihuc, Sarah, Wagner, Michael and Sonderegger, Morgan. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Mohammad, Saif M., Kiritchenko, Svetlana and Zhu, Xiaodan. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *ArXiv Preprint ArXiv:1308.6242*.

Mohammad, Saif M. and Turney, Peter D. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 26–34. LA, California: Association for Computational Linguistics.

Mohammad, Saif M. and Turney, Peter D. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Pang, Bo and Lee, Lillian. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Peng, Zeshan. 2017. *Acoustic feature-based sentiment analysis of call center data* (PhD Thesis). University of Missouri–Columbia.

Pereira, Moisés Henrique Ramos, Pádua, Flávio Luis Cardeal, Pereira, Adriano César Machado, Benevenuto, Fabrício and Dalip, Daniel Hasan. 2016. Fusing audio, textual, and visual features for sentiment analysis of news videos. In *Tenth International AAAI Conference on Web and Social Media*.

Pérez-Rosas, Verónica, Mihalcea, Rada and Morency, Louis-Philippe. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 973–982.

Plutchik, Robert. 1980. *Emotion: A psychoevolutionary synthesis*. Harper and Row.

Plutchik, Robert. 1997. The circumplex as a general model of the structure of emotions and personality. In R. Plutchik and H. R. Conte (eds.), *Circumplex models of personality and emotions*, 7–45. Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10261-001

Plutchik, Robert. 2000. *Emotions in the practice of psychotherapy: Clinical implications of affect theories*, Vol. 13. Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10366-000

Plutchik, Robert. 2001a. Integration, differentiation, and derivatives of emotion. *Evolution and Cognition*, *7*(2), 114–125.

Plutchik, Robert. 2001b. The nature of emotions. *American Scientist*, *89*(4), 344–350.

R Development Core Team. 2018. R: A language and environment for statistical computing (version 3.4.4) [computer software]. Vienna, Austria. https://doi.org/10.1511/2001.4.344

Razak, Aishah Abd, Abidin, Mohd Izani Zainal and Komiya, Ryoichi. 2003. Emotion pitch variation analysis in Malay and English voice samples. In *The 9th Asia-Pacific Conference on Communications 2003*, Vol. 1, 108–112.

Ribeiro, Filipe N., Araújo, Matheus, Gonçalves, Pollyanna, Gonçalves, Marcos André and Benevenuto, Fabrício. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 1–29. https://doi.org/10.1140/epjds/s13688-016-0085-1

Rosas, Verónica Pérez, Mihalcea, Rada and Morency, Louis-Philippe. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, *28*(3), 38–45. https://doi.org/10.1109/MIS.2013.9

Scherer, Klaus R. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165. https://doi.org/10.1037/0033-2909.99.2.143

Schuller, Björn, Batliner, Anton, Seppi, Dino, Steidl, Stefan, Vogt, Thurid, Wagner, Johannes, … Kessous, Loic. 2007. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Eighth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2007-612

Schuller, Björn, Steidl, Stefan and Batliner, Anton. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2009-103

Sheikh, Imran, Dumpala, Sri Harsha, Chakraborty, Rupayan and Kopparapu, Sunil Kumar. 2018. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 35–39. https://doi.org/10.18653/v1/W18-3305

Skinner, E. Ray. 1935. A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. *Communications Monographs*, *2*(1), 81–137. https://doi.org/10.1080/03637753509374833

Sobin, Christina and Alpert, Murray. 1999. Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research*, *28*(4): 347–365. https://doi.org/10.1023/A:1023237014909

Sperber, Dan and Wilson, Deirdre. 1986. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell.

Stolarski, Łukasz. 2018. Lack of effects of gender on the reading rate of long texts. *Sociolinguistic Studies*, 1*2*(3–4), 461–479. https://doi.org/10.1558/sols.32924

Stolarski, Łukasz. 2020. The influence of character's gender and the basic emotions of 'happiness' and 'sadness' on voice pitch in the reading of fiction. *Brno Studies in English*, *46*(1), 49–89. https://doi.org/10.5817/BSE2020-1-3

Stolarski, Łukasz. 2021. Comparison of key statistical instruments used in lexicon-based tools for sentiment analysis in the English language. *Token: A Journal of English Linguistics*, *13*: 219–248.

Taboada, Maite, Brooke, Julian, Tofiloski, Milan, Voll, Kimberly and Stede, Manfred. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2): 267–307. https://doi.org/10.1162/COLI_a_00049

Thelwall, Mike. 2017. Heart and soul: Sentiment strength detection in the social web with SentiStrength (summary book chapter). In J. Holyst (ed.), *Cyberemotions: Collective emotions in cyberspace*, 119–134. Berlin, Germany: Springer. https://doi.org/10.1007/978-3-319-43639-5_7

Thelwall, Mike and Buckley, Kevan. 2013. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the Association for Information Science and Technology*, *64*(8), 1608–1617. https://doi.org/10.1002/asi.22872

Thelwall, Mike, Buckley, Kevan and Paltoglou, Georgios. 2012. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology*, *63*(1), 163–173. https://doi.org/10.1002/asi.21662

Thelwall, Mike, Buckley, Kevan, Paltoglou, Georgios, Cai, Di, and Kappas, Arvid. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12): 2544–2558. https://doi.org/10.1002/asi.21416

Thelwall, Mike, Buckley, Kevan, Paltoglou, George, Skowron, Marcin, Garcia, David, Gobron, Stephane, … Holyst, Janusz A. 2013. Damping sentiment analysis in online communication: discussions, monologs and dialogs. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1–12). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-37256-8_1

Traunmüller, Hartmut and Eriksson, Anders. 1995. The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished Manuscript*.

Wallbott, Harald G. and Scherer, Klaus R. 1986. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, *51*(4), 690–699. https://doi.org/10.1037/0022-3514.51.4.690

Wöllmer, Martin, Weninger, Felix, Knaup, Tobias, Schuller, Björn, Sun, Congkai, Sagae, Kenji and Morency, Louis-Philippe. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, *28*(3), 46–53. https://doi.org/10.1109/MIS.2013.34

Wu, Wei, Zheng, Thomas Fang, Xu, Ming-Xing, and Bao, Huanjun. 2006. Study on speaker verification on emotional speech. In *Proceedings of Ninth International Conference on Spoken Language Processing, INTERSPEECH*, 2102–2105. Pittsburgh, Pennsylvania. https://doi.org/10.21437/Interspeech.2006-191

Zadeh, Amir, Chen, Minghai, Poria, Soujanya, Cambria, Erik, and Morency, Louis-Philippe. 2017. Tensor fusion network for multimodal sentiment analysis. *ArXiv Preprint ArXiv:1707.07250*. https://doi.org/10.18653/v1/D17-1115

Zhu, Xiaodan, Kiritchenko, Svetlana and Mohammad, Saif M. 2014. NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In *SemEval@ COLING*, 443–447. https://doi.org/10.3115/v1/S14-2077

Zuberbier, Erika. 1957. Zur Schreib-und Sprechmotorik der Depressiven. *Zeitschrift Für Psychotherapie Und Medizinische Psychologie*, *7*, 239–249.

Appendix Sentiment Scoring Validation

According to the results of the benchmark comparison presented in (Ribeiro, Araújo, Gonçalves, Gonçalves, and Benevenuto, 2016), VADER is one of the best sentiment analysis programs currently available. In the analysis they performed, 24 sentiment analysis tools were tested on 18 different validation datasets annotated by human raters. VADER was found to be the most consistent software for 3-level classification (positive vs. neutral vs. negative) and was ranked the most reliable of all the 24 tools. For 2-level classification (positive vs. negative) it also performed better than most other programs.

In another benchmark comparison summarized in Hutto and Gilbert (2014), VADER was compared to 11 other sentiment analysis tools. This time, the focus was on software offering numeric scoring. Hutto and Gilbert used validation materials which had been rated by human participants on a continuous scale. The correlation tests performed yielded varying results. VADER returned the most accurate scores in the dataset involving tweets pulled from Twitter. The correlation with the average human scores was 0.881, which was almost as good as the result for individual participants compared with these average scores (r=0.888). In other words, VADER performed at the level of the individual human scorers. In other datasets, the correlation coefficients were lower, although still higher in comparison to the results yielded by other sentiment analysis programs investigated.

Perhaps the most useful results for the purposes of the present paper are those found by Hutto and Gilbert (2014) while analysing New York Times opinion editorials validation dataset. The dataset includes 5,190 texts which are similar to the type of language analysed in Section 0. The correlation coefficient between the values produced by VADER and mean human scores was 0.492, which is visibly lower than the correlation measured for individual human raters in this dataset (r=0.745). Still, there are several ways in which VADER's performance may be improved. Firstly, one may exclude examples which are predictably problematic, such as texts including negation. As mentioned in Section 0, negation is an issue which has been discussed in sentiment analysis for a long time and only some software packages are capable of dealing with it. Nevertheless, the precision of scoring for such examples may be lower regardless of the solutions chosen. In order to investigate how VADER copes with negation, the present author wrote a script in Python which used VADER to score each text in the New York Times opinion editorials validation dataset. Additionally, each of these texts was classified according to whether or not it included negation. The correlation coefficient obtained for all examples was identical to the result reported in Hutto and Gilbert (2014) (r=0.492, n=5190, 95% CI: 0.472-0.513, p<0.0001), but slightly higher when texts with negation were excluded (r=0.521, n=4319, 95% CI: 0.499-0.542, p<0.0001). Secondly, the precision of VADER sentiment scoring may be improved by considering the categories of meaning specified in Section 0. For instance, in a test similar to the one described above, the additional

exclusion of questions slightly improved the correlation coefficient (0.521 → 0.527). This shows that the meaning classification used in this study could also be utilized in removing noise from the data. Thirdly, the accuracy of sentiment scores produced by VADER could be improved by focusing only on those values which are clearly positive or negative and ignoring more neutral results. This is evident when the New York Times opinion editorials validation dataset is narrowed down in this way. Namely, if any values between -1 standard deviation and +1 standard deviation are removed, the correlation coefficient for the remaining examples (which do not include texts with negation or questions) reaches 0.681 (n=1416, 95% CI: 0.652-0.708, p<0.0001). This result is not much below that obtained for individual human judges across the entire dataset (0.745). If the scope is further narrowed down to values outside -2 and +2 standard deviations, the correlation coefficient is even higher (r=0.801, n=181, 95% CI: 0.741-0.848, p<0.0001).

Finally, the accuracy of the sentiment scoring produced by VADER was also tested on a sample taken from "A Christmas Carol". The author randomly selected 108 text units from the first stave (or chapter) of the novella and graded their sentiment on a scale from -10 to +10. Next, with the aid of a Python script, the texts were graded in VADER. The correlation coefficients with the author's scores are remarkably similar to the results obtained for the New York Times opinion editorials validation dataset discussed earlier. For the whole sample, r=0.492 (n=108, 95% CI: 0.334-0.624, p<0.0001), while for the sample without negation and questions r=0.530 (n=98, 95% CI: 0.370-0.660, p<0.0001). Moreover, the exclusion of examples with sentiment scoring between -1 standard deviation and +1 standard deviation yields r=0.768 (n=28, 95% CI: 0.549-0.889, p<0.0001). All this suggests that VADER is a reliable sentiment analysis tool and with the appropriate selection of examples, its accuracy rate may approach the level of agreement achieved human scorers.