The usefulness of phonetically-motivated features for automatic laughter detection

Bogdan Ludusan
Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany

Abstract

The acoustic characteristics of laughter have been extensively investigated in speech sciences. However, despite the ability of several acoustic cues to discriminate between laughter and speech, the degree to which these are used in automatic systems varies. We analyze here the usefulness of four prosodic features: intensity, fundamental frequency, cepstral peak prominence (describing voice quality) and modulation spectrum (characterizing rhythm) for the automatic detection of laughter. We tested four machine learning algorithms on features extracted at the syllable level, considering also their derivative values. The results showed that using a few highly discriminative features returns a similar or better performance than that of systems employing a larger feature set or more complex learning algorithms. Moreover, a feature importance analysis revealed that the voice quality and rhythm-related measurements, less commonly used in automatic laughter detection systems, are better at discriminating between laughter and speech than traditional ones, such as intensity and fundamental frequency.

Index Terms: laughter, acoustic features, automatic detection, feature importance

1. Introduction

One of the laughter aspects extensively studied in speech sciences is its acoustic realization, also compared to the phonetic characteristics of speech [1]. This has been performed in terms of features describing both segmental (e.g., spectral) and prosodic (e.g., intensity, fundamental frequency) aspects. At the same time, there has been considerable interest in the speech technology community in the automatic detection of laughter from continuous speech (for a review on the topic, see [2]). Therefore, an apparent connection between the two fields exists, with studies coming from speech sciences complementing work conducted in speech technology.

Previous work in speech sciences found differences between laughter and speech with respect to several acoustic cues. One such feature is the signal intensity, with laughter exhibiting higher average intensity than speech [3] (but see [4, 5], conducted on smaller samples coming from fewer speakers, which found no difference in intensity between laughter and speech). Intensity, represented through different acoustic measures (e.g., root mean square energy, $0^{1/4}$ MFCC coefficient), has been widely used in laughter detection systems (e.g., [6, 7, 8]).

Another well-studied cue in connection with laughter is the fundamental frequency of the voice ($f_0$). Although significant variation with respect to $f_0$ has been seen between the various elements making up a laughter event [9], there seem to be consistent differences between laughter and speech in a number of $f_0$-related measures (mean and maximum value, $f_0$ range) [10, 11, 12, 5]. In all cases, the values of these measures were found to be higher for laughter than for speech. It is, thus, not surprising that a large number of automatic detection systems employed $f_0$ measures in their feature set (e.g., [7, 13, 14]).

There exists evidence supporting the separation between laughter and speech, based on voice quality characteristics, with laughter being produced either with a more pressed [15] or with a more breathy phonation [16, 3] than speech. The former was inferred from a formant analysis showing higher first formant values in laughter vowels, while the latter determined by means of a laryngoscopic study [16] and an analysis of acoustic voice quality features [3], respectively. Acoustic measures of voice quality have been used less often for automatic laughter detection [13, 17, 14], but have been included in a standard feature set employed also in laughter detection systems [18].

Considering the physiological aspects of laughter, a proposal was put forward by which laughter has its own distinct rhythm [19]. Further indications for this assumption were obtained through a perceptual experiment [20] and by investigating an acoustic-based parametrization or rhythm [21]. The later work showed that the particular rhythm representation employed may discriminate between laughter and speech at different modulation rates. Rhythm-based information has been used in several detection systems (e.g., [6, 7, 22]).

Laughter has been described also from a spectral point of view. However, the only observed differences compared to the speech spectrum were additional resonances around 1000 Hz [4]. The first three formants were found to be falling within the normal range of formants values [4], with most laughter vowels corresponding to central sounds (overlapping two or more vowel classes) [12, 15]. Despite the lack of evidence for the discrimination laughter/non-laughter based on spectral cues, these features have been widely used in automatic systems for laughter detection (e.g., MFCC [7] or PLP [13] coefficients).

We investigate here the usefulness of the features found to separate between laughter and speech for the automatic detection of laughter. We test several learning algorithms on a standard dataset, allowing for comparison with previous work. Moreover, we determine the importance of the examined features for laughter/speech discrimination. To our knowledge, no other study reported the importance of various prosodic features (but see [22] for the ranking of several intensity measurements for the discrimination of laughter and fillers from speech, [23] for spectral features and [24] for body movements).

2. Methods and materials

2.1. Corpus

The dataset employed in this study is part of the SSPNet Mobile Corpus [25], in which 60 pairs of speakers discussed, over the telephone, the usefulness of a number of items for surviving in

Abstract

The usefulness of phonetically-motivated features for automatic laughter detection

Bogdan Ludusan
Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany

Abstract

The acoustic characteristics of laughter have been extensively investigated in speech sciences. However, despite the ability of several acoustic cues to discriminate between laughter and speech, the degree to which these are used in automatic systems varies. We analyze here the usefulness of four prosodic features: intensity, fundamental frequency, cepstral peak prominence (describing voice quality) and modulation spectrum (characterizing rhythm) for the automatic detection of laughter. We tested four machine learning algorithms on features extracted at the syllable level, considering also their derivative values. The results showed that using a few highly discriminative features returns a similar or better performance than that of systems employing a larger feature set or more complex learning algorithms. Moreover, a feature importance analysis revealed that the voice quality and rhythm-related measurements, less commonly used in automatic laughter detection systems, are better at discriminating between laughter and speech than traditional ones, such as intensity and fundamental frequency.

Index Terms: laughter, acoustic features, automatic detection, feature importance

1. Introduction

One of the laughter aspects extensively studied in speech sciences is its acoustic realization, also compared to the phonetic characteristics of speech [1]. This has been performed in terms of features describing both segmental (e.g., spectral) and prosodic (e.g., intensity, fundamental frequency) aspects. At the same time, there has been considerable interest in the speech technology community in the automatic detection of laughter from continuous speech (for a review on the topic, see [2]). Therefore, an apparent connection between the two fields exists, with studies coming from speech sciences complementing work conducted in speech technology.

Previous work in speech sciences found differences between laughter and speech with respect to several acoustic cues. One such feature is the signal intensity, with laughter exhibiting higher average intensity than speech [3] (but see [4, 5], conducted on smaller samples coming from fewer speakers, which found no difference in intensity between laughter and speech). Intensity, represented through different acoustic measures (e.g., root mean square energy, $0^{1/4}$ MFCC coefficient), has been widely used in laughter detection systems (e.g., [6, 7, 8]).

Another well-studied cue in connection with laughter is the fundamental frequency of the voice ($f_0$). Although significant variation with respect to $f_0$ has been seen between the various elements making up a laughter event [9], there seem to be consistent differences between laughter and speech in a number of $f_0$-related measures (mean and maximum value, $f_0$ range) [10, 11, 12, 5]. In all cases, the values of these measures were found to be higher for laughter than for speech. It is, thus, not surprising that a large number of automatic detection systems employed $f_0$ measures in their feature set (e.g., [7, 13, 14]).

There exists evidence supporting the separation between laughter and speech, based on voice quality characteristics, with laughter being produced either with a more pressed [15] or with a more breathy phonation [16, 3] than speech. The former was inferred from a formant analysis showing higher first formant values in laughter vowels, while the latter determined by means of a laryngoscopic study [16] and an analysis of acoustic voice quality features [3], respectively. Acoustic measures of voice quality have been used less often for automatic laughter detection [13, 17, 14], but have been included in a standard feature set employed also in laughter detection systems [18].

Considering the physiological aspects of laughter, a proposal was put forward by which laughter has its own distinct rhythm [19]. Further indications for this assumption were obtained through a perceptual experiment [20] and by investigating an acoustic-based parametrization or rhythm [21]. The later work showed that the particular rhythm representation employed may discriminate between laughter and speech at different modulation rates. Rhythm-based information has been used in several detection systems (e.g., [6, 7, 22]).

Laughter has been described also from a spectral point of view. However, the only observed differences compared to the speech spectrum were additional resonances around 1000 Hz [4]. The first three formants were found to be falling within the normal range of formants values [4], with most laughter vowels corresponding to central sounds (overlapping two or more vowel classes) [12, 15]. Despite the lack of evidence for the discrimination laughter/non-laughter based on spectral cues, these features have been widely used in automatic systems for laughter detection (e.g., MFCC [7] or PLP [13] coefficients).

We investigate here the usefulness of the features found to separate between laughter and speech for the automatic detection of laughter. We test several learning algorithms on a standard dataset, allowing for comparison with previous work. Moreover, we determine the importance of the examined features for laughter/speech discrimination. To our knowledge, no other study reported the importance of various prosodic features (but see [22] for the ranking of several intensity measurements for the discrimination of laughter and fillers from speech, [23] for spectral features and [24] for body movements).

2. Methods and materials

2.1. Corpus

The dataset employed in this study is part of the SSPNet Mobile Corpus [25], in which 60 pairs of speakers discussed, over the telephone, the usefulness of a number of items for surviving in
a polar region. The speakers were English native speakers (63 females, 57 males), aged 18 to 64 years old, and did not know each other. A total of 2763 11-second clips were extracted from these recordings, and have been previously used in a social signals challenge [26]. The resulting dataset consists of about 8.5 hours of speech materials. It contains annotations for laughter and fillers, the former being employed in this study. A total of 1158 laughter events were identified in the recordings.

2.2. Feature extraction

We tested the usefulness of the following features for automatic laughter detection:

- root mean square energy (en): a parametrization of the intensity of the speech signal;
- fundamental frequency (f0): the acoustic correlate of pitch;
- cepstral peak prominence (cpp): the amplitude of the cepstral peak relative to the regression line over the entire cepstrum [27], a voice quality measure (lower values corresponding to a more breathy voice);
- modulation spectrum (ms): linked to rhythm, encoding the variation of the temporal envelope of the signal; the employed implementation represents a ratio between the intensity of the signal and the noise [28].

The energy of the signal, the fundamental frequency and the cepstral peak prominence were computed by means of the VoiceSauce software [29], using an analysis frame of 25 ms, a frame shift of 10 ms and the default values for the remaining parameters. For determining f0, the Straight algorithm [30] was chosen. The modulation spectrum was obtained by means of the AM_FM_Spectra toolbox [31], using a window of 1.5 s and a frame shift of 50 ms. We then obtained the difference between the modulation spectrum of the laughter and speech classes, based on the DUEL corpus [32], by applying the process proposed in [33]. The elements of the resulting difference matrix that best separated the laughter and speech classes were identified and the average value of these elements was considered to be our modulation spectrum feature value.

All four features characterize phenomena described at a level higher than the frame. Therefore, we employed an approach considering an analysis level equivalent to that of the syllable, similar to [14, 22]. We segmented the recordings into syllable-like units using a previously proposed sonority-based system [34]. The posterior probabilities returned by a broad phonetic class recognizer trained on 100 hours of English recordings were used to compute a sonority function. The value of the function was computed for each frame, by multiplying the probability of each phonetic class in that frame with the sonority of that class (ranging from 1, for plosives, to 7, for vowels; silence had a sonority value of 0). We also determined the voice activity regions, employing a state-of-the-art system [35]. These were then employed as a mask for the sonority output, by setting to 0 all frames which were found not to have voice activity. The resulting function was smoothed with a Gaussian filter with $\sigma = 1$ and the valleys between two consecutive peaks of the function were considered syllable boundary candidates. Next, we merged any syllable shorter than 50 ms with its neighbour (in order for each syllable-like unit to have a value for the ms feature).

We considered in the analysis only the parts of the recordings that were divided into syllables (the parts marked by the voice activity detector or the broad phonetic class recognizer as silence were discarded). For each syllable-like unit, we computed the value of each feature by taking the mean value over all the frames that made up the unit. This gave us a set containing 4 prosodic features for each analysis unit. Additionally, as many speech technology systems take into account also the variation of the features between consecutive analysis units, we tested here feature sets that included the first ($\Delta$) and the second order derivatives ($\Delta\Delta$) of the four prosodic cues, as well. Thus, a second set was composed of 8 values per unit (prosodic + $\Delta$ features), while a third one had 12 values (prosodic + $\Delta + \Delta\Delta$ features).

2.3. Experimental setup

We tested several learning algorithms for discriminating between laughter/non-laughter segments: Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF) and Neural Networks (NN). We employed, for all of them, the implementation offered by the scikit-learn library [36, ver 1.0.2]. All experiments were run in a train/test setting, similarly to the ComParE challenge [26]. We employed the challenge train and the development sets (90 speakers, in total) for performing parameter search and training the final model, while the test set (the remaining 30 speakers) was used for validation. Thus, the tests were done in a speaker-independent fashion. For parameter optimization, we performed grid search cross-validation, by varying the following parameters: for LR, the solver, the class weight, the penalty type and the parameter C, representing the inverse of the regularization strength. For RF, we varied the number of trees, the split criterion function, the maximum depth of the tree, the minimum number of samples to split a node, the minimum number of samples in a leaf node and the number of features to consider for splitting. Lastly, for NN we tested several network architectures with two and three hidden layers, varying also the activation function, the initial learning rate and the exponential decay rates (parameters $\beta_1$ and $\beta_2$).

We evaluated the classification performance by comparing the labels returned by the systems with those of the reference segments. The reference segments were determined as follows: if a syllable-like unit was overlapping with an interval marked in the annotations as being laughter, it was considered to be laughter, otherwise speech. The reference segments part of the train and dev set were used for training the models. As the data was highly imbalanced (a proportion of 1:21 laughter segments to non-laughter segments) and our focus was the less frequent class (laughter), we chose metrics that evaluated the performance of the system on this class. Therefore, we computed precision (number of correctly classified laughter segments out of the total number of segments), recall (number of correctly classified laughter segments out of the total number of reference laughter segments) and F1-score (the harmonic average between precision and recall). Moreover, we varied the decision threshold and computed the precision and recall values for each threshold value, determining from them the area under the precision-recall curve (PRC). Since we compared our results to those of the systems proposed as part of the ComParE challenge, we adopted the evaluation metric employed there, the area under the receiver operating characteristic curve, as well.

We were also interested to find out which of the investigated features are more discriminative between laughter and non-laughter. For this, we performed a feature importance permutation analysis. First, the performance of the system with the original feature set was computed. We used here as evaluation measure of the system the area under the receiver operat-
Table 1: Laughter detection results for the different algorithms (Naive Bayes - NB, Logistic Regression LR, Random Forest - RF, and Neural Networks - NN) and features sets tested here. The feature sets include: prosodic features, prosodic features and their deltas, as well as prosodic features, deltas and double deltas. The precision, recall, F1-score, area under the receiver operating characteristic curve (ROC) and area under the precision-recall curve (PRC) are reported in the table. The bold values represent the best attained performance for that particular metric.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Prosodic features</th>
<th>Prosodic features + Δ</th>
<th>Prosodic features + Δ + ΔΔ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>.349</td>
<td>.350</td>
<td>.349</td>
</tr>
<tr>
<td>NN</td>
<td>.536</td>
<td>.235</td>
<td>.326</td>
</tr>
</tbody>
</table>

3. Results

3.1. Laughter detection

The results obtained with the three feature sets, employing the different machine learning algorithms, are illustrated in Table 1. We can see differences between the different learning paradigms employed here, with the NB system having a rather balanced precision and recall, the LR algorithm strongly favouring recall over precision and the other two systems having a high precision, at the expense of a lower recall. There seem to be some differences in overall performance between the four learning algorithms, as well as smaller differences between the three feature sets (adding derivative information usually helped classification, although not for NB).

In order to have a more similar evaluation to those reported in previous studies, we evaluated the performance of the learning algorithms also at the frame level. This was done by expanding each syllable-like segment into a number of frames corresponding to its length in seconds multiplied by 100 (equivalent to having a 10 ms shift between frames) and assigning the label that was given by the system to that segment to every frame. This was then compared against the reference frame-based annotation, derived similarly, from the gold standard syllable-level segments. The frame-level evaluations of the systems having the highest precision, recall and F1-score (marked with bold fonts in Table 1) are presented in Table 2. We can see a similar or slightly higher performance in terms of precision, recall, F1-score and PRC measures, and a considerable increase in area under the ROC curve for the frame-level evaluation, compared to the syllable-level one.

3.2. Feature importance

The feature importance, as given by the permutation analysis, is illustrated in Figure 1. Looking at the left panel of the figure, displaying the importance of the four prosodic cues, the highest importance was observed for $ms$, followed by $cpp$, $f0$ and the lowest for $en$. This ranking changes slightly when the delta or delta and double delta features are considered, but is rather consistent in the two cases: $cpp$ is the most discriminative feature, followed by $ms$ and, lastly, by $en$ and $f0$.

The delta features have a lower importance (especially those of $ms$ and $f0$), except for the derivative of $cpp$, which exhibits a similar importance to the fundamental frequency of the signal. A similar picture can be seen also for the double delta features, with $\Delta cpp$ and $\Delta en$ having the highest importance (even higher than that of $en$ and $f0$) and $\Delta ms$ having the lowest one among the four.

4. Discussion and conclusions

Comparing our findings to those of other systems employing syllable-level decision and an identical evaluation setting to the one we used here [14, 22], we notice a higher area under the ROC curve value in our case (.885, compared to .846 and .859, respectively). We acknowledge that other laughter detection proposals part of the ComParE challenge performed better than our system on the same test set (reporting ROC values higher than 0.9). Nevertheless, one needs to take into account that we did not evaluate the parts of the corpus that were found to be silence, resulting in a less imbalanced laughter/speech dataset than in their case. We have seen, by comparing the results at the segment-level with those at the frame-level, that an evaluation in terms of area under the ROC curve will result in a better performance when the class imbalance widens. For this reason, it is recommended in the literature to consider other evaluation measures, such as the area under the precision-recall curve, when dealing with imbalanced data [37].
Among the many automatic laughter detection approaches that employed the SSPNet corpus, only few of them focused their system evaluation on the laughter class [8]. Our highest recall system had similar performance to their highest recall system, while our highest precision system had a lower precision than their best precision system. Lastly, in terms of overall F1-score, we observed a better performance here. [8] used spectral features along with speech energy and f0 information, which suggests that similar or better performance may be obtained when more relevant features are employed instead of generic ones (MFCC, in this case), even if no deep learning paradigm is employed for the classification system. However, using prosodic features (voice quality and rhythm-related) instead of MFCCs is not the only difference between these approaches. Different granularities were considered for the decision laughter/non-laughter: syllable-like (in order to better integrate the long-term information encoded in our features), vs. frame-based decisions, and also the experimental setting differed between studies: a train/test paradigm here vs. a 6-fold cross-validation in [8]. Further investigations will be necessary to determine whether any seen advantage is due to the nature of the features (prosodic, long-term) or the level at which the decision is taken (syllable-like).

The feature importance analysis revealed interesting results: the voice quality and rhythm-related features were ranked higher than the fundamental frequency and speech intensity features. This is rather surprising considering that the latter two features are more often used in laughter detection systems. This tendency, favouring the use of the signal energy and f0 may be due to the historic use of these features in frame-based speech technology applications. These unexpected findings may be explained by the level of analysis normally employed in detection systems (the frame), which might be less well-suited for some of the features we considered, such as ms. This level mismatch might explain the findings in [6], where no performance improvement was seen when adding modulation spectrum information to spectral features. Another possible explanation for these results may be the different parametrizations for voice quality and rhythm employed here. For instance, cpp has been found to be the best acoustic correlate for voice quality in continuous speech [38], while the measures used in previous work (jitter/shimmer [13], spectral tilt [14, 17]) were ranked consistently lower.

With regards to the delta and double-delta features, we see that they play a substantially less important role than the features from which they were derived. This is especially the case for the derivatives of ms, but also for those of f0. In contrast, the derivatives of cpp and, partially those of en, have a similar if not higher importance than that of the energy of the signal or its fundamental frequency, which suggests they may interact with other features for the discrimination laughter/non-laughter. These findings seem to be in contrast to those of [22] (where also syllabic-level features were employed), in which intensity and derivatives of intensity features were ranked the highest. However, we used different measures and/or parametrizations besides intensity, compared to [22], which might have a higher discriminative power, and may explain these discrepancies.

To summarize, we have seen that using exclusively features which have been previously shown to discriminate between laughter and non-laughter, in conjunction with simple learning architectures, brings a similar or higher performance to using generic features and/or more complex machine learning systems. Moreover, the results showed that less frequently used prosodic characteristics, voice quality and rhythm-related ones, discriminate better than more established measures for laughter detection, speech signal intensity and fundamental frequency. While the current study made use of the SSPNet corpus, a standard dataset for automatic laughter detection, we would like to confirm our findings also on other corpora. Further work may explore the use of the prosodic representations employed here in conjunction with spectral information, to better understand their importance in discriminating between laughter and speech.

5. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 461442180.

6. References


