Towards a Level Assessment System of Amusement in Speech Signals: Amused Speech Components Classification

Kevin El Haddad, Huseyin Çakmak, Stéphane Dupont, Thierry Dutoit
TCTS lab - University of Mons, Belgium

ABSTRACT
In this paper, we expose our work on classification of smiled vowels, shaking vowels, and laughter syllables. This work is part of a larger framework that aims at assessing the level of amusement in speech using only audio cues. Indeed all of these three categories occur in amused speech and are considered to express a different level of amusement. Four novel features are used to accomplish this task. With only those four features, we are able to obtain good classification results with different systems. Among the compared systems, the best one achieved 21.04% error rate, therefore an accuracy well above chance.

Index Terms— classification, amused speech, speech-smile, laughter, acoustic features.

1. INTRODUCTION

Emotion recognition is currently one of the hottest research topics due to its potential in many different application areas. Emotion can trigger expressions in different modalities, one of the most important being speech. Most previous works on emotion recognition in speech focused on classifying several types of emotions [1, 2, 3]. Our work in this paper rather focuses on an emotion of positive valence: amusement. We aim at assessing the arousal level of amusement in speech using information from the audio modality only. We propose a system able to classify speech-smile vowels, shaking vowels, and laughter syllables. There are two main components of amused speech. The first one is the smile which is an audibly identifiable component [4, 5]. The second one is laughter. Laughter interrupting and/or intermingling with speech causes what is called speech-laughs [6].

The estimation of the amusement arousal level in speech needs, by definition, the establishment of different levels. These levels should therefore be based on the components that constitute amused speech, i.e., speech-smile and laughter (and/or speech-laughs). Indeed, our hypothesis is that the arousal level of amusement in a sentence depends on the presence of these two components and is correlated with their intensity. Smiled vowels, shaking vowels, and laughter syllables can be found in amused speech and could each be representative of a level. Therefore their classification after being detected in a sentence would contribute to an accurate arousal level assessment. For instance, the shaking vowels will be considered in this work as low level speech-laughs (see Section 2). Therefore, the detected shaking vowels should correspond to a level lower than the one to which the detected laughter corresponds.

Emotion arousal dimension estimation and prediction have been addressed in previous work [7, 8, 9, 10, 11].

In what follows, we will first give a detailed description of our three different classes, an overview of the datasets from which they were extracted and the extraction procedure in Section 2. Section 3 contains a description of the features extracted from the data and used for classification. In Section 4, we compare several systems for this classification task: a k-Nearest Neighbor (kNN), Support Vector Machine with Linear (SVM-Lin) and Polynomial (SVM-Poly) kernels and a one hidden layer feedforward neural network (NN). Results will be given and discussed in Section 5. Finally we will conclude in Sections 6.

2. DATA

This section will first describe in more details each of the three different classes we are working with. It will then present the datasets that are used and will finally give details about the data extraction procedure.

2.1. Description

Speech-smile is a term used to describe the alteration of speech due to smiling. Smiles can indeed be audibly identifiable in speech [4, 5].

Laughter syllables represent the laughter bursts that are inserted in speech during a speech-laugh. A prototypical laughter event is a sequence of fricatives and vowels. A laughter syllable is the succession of a fricative and a vowel (e.g., a “ha” sound). Please note that we use the term “fricative” here to refer only to the /h/ which is the most present
a sound proof room with head mounted Sennheiser HSP4 amused but without laughing. These data were recorded in an actor was asked to read sentences in French while sounding finally to improve the systems presented in [14, 15, 16]. The Belgian professional actor. This dataset’s purpose was originally to improve the systems presented in [14, 15, 16].

The data were selected from three different datasets. The first one (DS1) contains speech-smiles recorded in French from a one (DS1) contain speech-smiles recorded in French from a Belgian professional actor. "Laughing vowels" refers to laughter bursts occurring in a sustained vowel as described in [14]. Indeed, the two speakers were asked to sustain the pronunciation of a vowel while watching funny videos. Laughter eventually interrupted the vowels being sustained. The data from speaker 1 was recorded in the same conditions as DS1 while the recording of speaker 2 was made in a quiet room with a Rode Podcaster microphone at a sampling frequency of 48 kHz and stored in a 16 bit PCM WAV format.

The data were all downsampled to a sampling frequency of 16 kHz and stored in a 16 bit PCM WAV format.

Table 1 gives a summary of the datasets used here. It contains the language (Lang), the recording conditions (Rec Cond.) of the datasets and what type of data was extracted from it (Sm = smiled vowels, Sh = shaking vowels, Laugh = laughter syllables). The recording conditions columns contain whether the data recorded are “clean” or “noisy” and whether they are naturally expressed or acted.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lang</th>
<th>Rec. Cond</th>
<th>Sm</th>
<th>Sh</th>
<th>Laugh</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>French</td>
<td>clean/acted</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DS2</td>
<td>English</td>
<td>noisy/natural</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>DS3</td>
<td>paralinguistic</td>
<td>clean/natural</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Dataset summary table: In the Lang column, "paralinguistic" means that there was no precise language recorded, since the dataset does not contain words (although the data come from French speaking persons). The + indicates the presence of the type while the - indicates its absence.

2.2. Database

The data were selected from three different datasets. The first one (DS1) contain speech-smiles recorded in French from a Belgian professional actor. This dataset’s purpose was originally to improve the systems presented in [14, 15, 16]. The actor was asked to read sentences in French while sounding amused but without laughing. These data were recorded in a sound proof room with head mounted Sennheiser HSP4 microphone and Rode Podcaster microphone at 48 kHz of sampling frequency both and stored in a 16 bit PCM WAV format. In this work only the data recorded with the HSP4 head mounted microphone will be used. The second dataset is a subset of the ICSI Meeting corpus database [17] which contains spontaneous recorded meeting conversations. Annotations of non-verbal vocal sounds are also provided with this dataset such as laughter or laughter occurring together with speech. The laughter segments occurring with speech (speech-laughs) were segmented and only the ones in which the speech and laughs came from the same speaker were considered. Thus forming the dataset DS2. The third dataset (DS3) contains laughing vowels recorded from two different French speakers, one of them (speaker 1) being the same person as in DS1. The other person (speaker 2) being a naive actor. "Laughing vowels" refers to laughter bursts occurring in a sustained vowel as described in [14]. Indeed, the two speakers were asked to sustain the pronunciation of a vowel while watching funny videos. Laughter eventually interrupted the vowels being sustained. The data from speaker 1 was recorded in the same conditions as DS1 while the recording of speaker 2 was made in a quiet room with a Rode Podcaster microphone at a sampling frequency of 48 kHz and stored in a 16 bit PCM WAV format.

The data were all downsampled to a sampling frequency of 16 kHz and stored in a 16 bit PCM WAV format.

The data were all downsampled to a sampling frequency of 16 kHz and stored in a 16 bit PCM WAV format.

2.2. Database

The data were selected from three different datasets. The first one (DS1) contain speech-smiles recorded in French from a Belgian professional actor. This dataset’s purpose was originally to improve the systems presented in [14, 15, 16]. The actor was asked to read sentences in French while sounding amused but without laughing. These data were recorded in a sound proof room with head mounted Sennheiser HSP4
alignment. The obtained annotations were then checked manually. A total of 262 and 73 smiled vowels were extracted from DS1 and DS2.

The laughter syllables were extracted from DS3, thus yielding to a total of 1004 samples. These were extracted from laughing vowels occurring in the vowels /a/, /i/, /o/, and /e/ for speaker 1 and in the vowels /ʌ/, /o/, /ɪ/, /ɛ/ and /e/ for speaker 2. The extraction of the laughter syllables was based on annotations made manually. Each fricative and each vowel were annotated separately, then each fricative was extracted with the vowel or fricative that followed it.

Concerning the shaking vowels, a total of 17 and 31 instances were extracted from DS1 and DS2 respectively. Indeed when reading the sentences during the recordings of DS1, the actor, focused on his instructions of having to sound amused, also produced shaking vowels and speech laughs in some of his utterances. Shaking vowels were also extracted from the speech-laughs of DS2 in which they occurred occasionally. The extractions were also based on annotations made manually which tried to delimit them in the sentence the most accurately possible. We thus gathered 48 shaking vowel samples in total.

3. EXTRACTED FEATURES

In our previous work [18], different sets of features were studied for the speech-smile/shaking vowels classification task. In that study, we introduced Stability-based Features (SF) under the hypothesis that some shaking vowels characteristics (such as the pitch and the signal power) might be less stable than the smiled vowels ones. This study showed that the combination of the SF combined with temporal and spectral features gave the best results (better than combinations containing Mel-Frequency Cepstrum Coefficients (MFCC)) with a kNN. In what follows we describe the SF as well as the other features used in this article.

3.1. Stability-Based Features

The Stability-based Features are based on the observations made in Section 2. They will also be combined to MFCC and pitch based acoustic features which are commonly used features in speech and emotion recognition. The different sets of features obtained will be compared for training different systems.

These statistical features are based on the stability of the pitch and signal power values. We consider the speech-smile vowels to have properties such as pitch and power that are more stable than the shaking vowels ones. This is due to the fact that the latter are formed by a "concatenation" of vowels between which air pulses might interfere. Since the laughter syllables are formed by a fricative followed by either a vowel or another fricative, we also expect these parameters stability to be affected by this pattern.

So, in order to represent this stability, the pitch is first estimated on each sample of smile vowel, shaking vowel and laughter syllable using the ESPS method of the Snack library [19]. This is done on a sliding window of length 20 ms and shifted by 10 ms. The derivative of the obtained pitch is also calculated. Finally the standard deviation of the pitch derivative and of the residuals of a linear regression fitted on the pitch values are computed to form the first 2 features. The standard deviation values turned to have a skewed distribution and a log transformation of these data was necessary to obtain better discriminating features.

Then, the log-power envelope of the signal is considered. During our analysis this value showed a discriminating pattern for the three classes. In fact, it showed downward peaks in shaking vowels at the vowels separation. It also showed higher values for the vowel parts and lower ones for the fricative parts of the laughter vowels. Compared to these behaviors, the smiled vowels log-power envelope variation seemed to be monotonous.

The log-power of the signal is computed using the following formula:

\[ P(i) = 10 \log \frac{x(i)^2}{\Delta T} \]

\[ x(i) \text{ being the } i^{th} \text{ signal sample and } \Delta T \text{ the sampling period.} \]

From this log-power value, we estimate the envelope by keeping the maximum values of a 10 ms frame shifted by 10 ms.

Then, the same approach used for the pitch is used for the power envelope, giving us the last 2 features for this set.

Before computing these features, 15% of the beginning and end of each segment were removed so that the transitions with the preceding and following phonemes affect less the extracted features.

This set of features will be referred to as "Stability-based Features (SF)" in the remainder of this article.

3.2. Commonly Used Features

As mentioned in Section 3.1, the SF will be used with features that are commonly used in speech and emotion recognition. These features are based on the MFCCs and on the pitch estimation. The features created from the pitch are the mean and standard deviation values of the pitch and pitch derivative values. Thus forming a feature vector with four elements. This vector will be referred to as "MFCC" in the following.
4. SYSTEMS COMPARISON

4.1. Overview

The sets of features presented in Section 3 were used to train several systems in order to compare their efficiency. The systems compared here are kNN, SVM with a linear and polynomial kernels, and a one hidden layer feedforward neural network. These systems were chosen for their efficiency in classification tasks and their ability to handle a small number of data. The pipeline shown in Fig 2 is applied to all the previously mentioned systems.

Considering the difference in the number of samples available for each class (48 shaking vowel samples, 335 smiled vowels, and 1004 laughter syllables), a first random sampling is applied on the smiled vowel and the laughter syllables to obtain a balanced dataset. Thus, 50 samples of each of these classes are randomly selected and gathered with the 48 shaking vowel samples. We obtain a new balanced dataset with a total of 148 samples. This new dataset is then randomly split into 75% of it to train the system and 25% of it to test it. During the training step, a k-fold cross validation scheme is undertook to tune the system’s set of parameters. When the optimal set of parameter is found, it is used to train the system using the whole training data this time. The testing data are then used to evaluate the final obtained system and an Error Rate (ER) is computed. This is the mistakenly classified sample in the test set to the total number of samples in this set.

The entire process is repeated 1000 times, thus obtaining 1000 ER values for each system. The systems ER distributions will be compared to each other.

4.2. Implementation Details

As mentioned previously, in this article we compare the efficiency of a kNN, SVMs with linear and polynomial kernels and a one hidden layer feedforward neural network. As explained in Section 4.1, the process described in Fig 2 was applied to all the systems. This section will give more details about our implementations of the systems.

First, in the k-folds cross-validation step, the number of folds was chosen to be 4 so that the folds contain 25% of the training data each.

A list of the kNN, SVM-Lin, SVM-Poly and NN’s tuned parameters are given in the Table 2.

<table>
<thead>
<tr>
<th>System</th>
<th>kNN</th>
<th>SVM-Lin</th>
<th>SVM-Poly</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>k</td>
<td>C</td>
<td>degree,C,γ</td>
<td>n,d</td>
</tr>
</tbody>
</table>

Table 2. Tuned parameters per system. k: number of neighbors, C: cost parameter, degree: polynomial degree, γ: scale parameter, n: number nodes in the hidden layer, d: weight decay

The values of these parameters varied for every iteration of the process described in Fig 2.

Concerning the neural network, we chose a small architecture taking into account the small number of data available. It is composed of an input layer with 4 nodes (since we have 4 input features), an output layer with 3 nodes (3 classes) and only one hidden layer. The optimal number of nodes in the hidden layer is found in the parameter tuning step. The training algorithm is the standard gradient descent backpropagation algorithm with a weight decay term. The optimal weight decay term is also found in the parameter tuning step.

5. RESULTS

Table 3 shows the mean values of the ER distribution per system.
<table>
<thead>
<tr>
<th>System</th>
<th>SF</th>
<th>MFCC</th>
<th>f0</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>28.2%</td>
<td>31%</td>
<td>33.9%</td>
</tr>
<tr>
<td>SVM-Lin</td>
<td>25.3%</td>
<td>31.2%</td>
<td>38.1%</td>
</tr>
<tr>
<td>SVM-Poly</td>
<td>24.4%</td>
<td>29.1%</td>
<td>30.8%</td>
</tr>
<tr>
<td>NN</td>
<td>23.8%</td>
<td>30.4%</td>
<td>28.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>SF+MFCC</th>
<th>SF+f0</th>
<th>MFCC+f0</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>kNN</td>
<td>27.8%</td>
<td>25.9%</td>
<td>28.3%</td>
<td>26.9%</td>
</tr>
<tr>
<td>SVM-Lin</td>
<td>27.07%</td>
<td>23.1%</td>
<td>28.7%</td>
<td>25.2%</td>
</tr>
<tr>
<td>SVM-Poly</td>
<td>27.7%</td>
<td>21.04%</td>
<td>27.4%</td>
<td>27.6%</td>
</tr>
<tr>
<td>NN</td>
<td>26%</td>
<td>21.9%</td>
<td>27.9%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Table 3. Mean error rate results of each system per feature set. The systems results when using the features without combining them are in the upper table while the ones using features combinations are in the lower table. The best result in each table are statistically significant results under a 95% CI Student’s t-test (in bold).

A first observation of the obtained results shows that the SF feature vector outperform the MFCC and the f0 with respect to any system. Also, we can note that when combining the feature vectors (lower table), the combination of the SF and f0 feature vectors (SF+f0) is more efficient than any other combination with respect to any system.

When comparing the systems, the SVM with polynomial kernel obtained the best score when used with the SF+f0 combination (21.04%). Considering only one feature vector (higher table) the best score was obtained by the neural network when used with the SF feature vector (23.8%).

A 95% Confidence Interval (CI) Student’s t-test was ran on the obtained ER distributions in order to check if the differences between the obtained scores are significant. The majority of the differences appeared to be significant (690 out of 756 are significantly different). The best results obtained in each table proved to be significant with respect to all other results in both tables.

We first conclude that using the SF alone gives better results than using the MFCC or the f0 vectors alone with any of the systems presented here. When combining the features, the results show that the combination SF+f0 gave better results than all other feature sets (combined and not combined) when also used in any system compared here. The SVM with polynomial kernel used with the SF+f0 combination gave statistically significant better results than all other systems and the single layer neural network’s results are significantly better than any other system when using non-combined feature vectors.

6. CONCLUSION

A smiled vowels, shaking vowels and laughter syllables classification study was presented in this article. This work will serve in amusement arousal level assessment in speech which aims in future work to correspond each class to a certain level of amusement.

To serve our purpose in this work, we used novel stability-based features (SF) combined with commonly used features in the speech and emotion recognition areas. These combinations of features were used with different systems for comparison purpose.

The results presented in this paper first showed that, when comparing the SF to the MFCC and the f0 vectors without combinations, this feature vector gave better results when used with any system presented here (kNN, SVM-Lin, SVM-Poly and NN). With non-combined vectors, the single-layer neural network gave significantly better results than the other systems when used with the SF feature vector. Also, it has been shown that the combination of SF and f0 gave the best results in general. In particular, an SVM with polynomial kernel used with a combination of SF and f0 vectors gave significantly better results than any other system.

Further work will focus on collecting a larger database and finding other features (and/or feature combinations) in order to ameliorate our classification system. An interesting perspective is to test the sets of features presented here with systems for the purpose of automatic detection of the three classes mentioned in this paper. This will probably be a frame wise automatic detection system. Also, in future work, more classes could be defined based on other more precise components of amused speech (level of smiling, intensity of laughter syllables, etc.).
REFERENCES


