Amused Speech Components Analysis and Classification: Towards an Amusement Arousal Level Assessment System

Kevin El Haddad, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit

TCTS lab - University of Mons

Abstract

In this paper, we present our work on analysis and classification of smiled vowels, chuckling (or shaking) vowels and laughter syllables. This work is part of a larger framework that aims at assessing the level of amusement in speech using the audio modality only. Indeed all of these three categories occur in amused speech and are considered to contribute in the expression of different levels of amusement. We first analyze these three amused speech components on the acoustic level. Then, we improve a classification system we previously developed. With a limited amount of data and features, we are able to obtain good classification results with different systems. Among the compared systems, the best one achieved 82.8% of accuracy, therefore outperforming chance.

Keywords: `Amusement Intensity Level`, `Laughter`, `Smile`, `Paralinguistic`, `Affective Computing`, `Machine Learning Application`

1. Introduction

Affective computing and more specifically, emotion recognition is currently one of the hottest research topics due to its potential in many different application areas. Applications involve Human-Computer Interactions (HCI), medical and social areas. Emotions can trigger expressions in different modalities, one of the most important being speech. Previous work on emotion recognition from speech mostly focused on classifying several types of emotions [1, 2, 3].

The work we present here rather focuses on a single emotion of positive valence: amusement. It is part of a larger framework of assessing accurately the amusement arousal level in speech using information from the audio modality only. In fact most of the previous work related the emotion arousal or intensity level estimation tackles the problem by using multimodal data. For instance Patwardhan and Knapp present work in [4] on estimating the intensity level of anger expressions using speech and motion capture data. Also, Yoshiko et. al. estimate anger intensity level using speech and linguistic data [5]. Dhall and Goecke propose an estimation of different levels of smile and laugh using also multimodal data in [6] (smiling and laughter in that work aren't necessarily expressions for amusement).

In this paper, we propose a preliminary analysis and classification work. Indeed, we present analyses of amused speech components and improved results on classifying these components compared to a previous classification attempt [7]. The ultimate goal being, as mentioned earlier, to estimate accurately the intensity level of amusement in speech using a single modality only, the audio.

Smiled vowels, chuckling (or shaking) vowels [8] and laughter syllables can all be found in amused speech and are therefore considered here as its components. We will refer to them as Amused Speech Components (ASC). We consider that two main dimensions constitute amusement in speech. The first one is the smile which is not only a visual expression, but also identifiable audibly [9, 10]. The second one is laughter. Laughter interrupting and/or intermingling with speech causes what is called speech-laughs [11]. The estimation of the amusement arousal/intensity level in speech needs, by definition, the establishment of different levels. These levels should be based on the two main amused speech components previously mentioned, i.e. speech-smile and laughter (and/or speech-laugh). Indeed, our hypothesis is that amusement intensity level of an uttered sentence depends on the presence of these two components and is correlated with their intensities. The presence (or absence) of each ASC, and their combination in an uttered sentence, could be representative of an intensity level.

In our previous works regarding ASC [12, 7], classification features were extracted based on observations made on data we collected. The efficiency of these features was tested in a classification task with different machine learning algorithms. Motivated by the good results obtained in these works, we present a more detailed analysis of the ASC, new discriminating feature sets, and improved classification results using these features.

The remaining of this paper is organized as follows. We first present the data collected for the purpose of this work in Section 2. In that section, we will start by giving a more detailed definition of the ASC. We will then present the data collection protocol, followed by the analysis of data. Section 3 will summarize our previous works and recall the initial ASC classification results obtained. Then, the new feature sets will be introduced in Section 4. The improved classification results obtained will be presented in Section 5. We will finally conclude and give our perspectives for future work in Section 6.

2. Amused Speech Components Data

2.1. Description

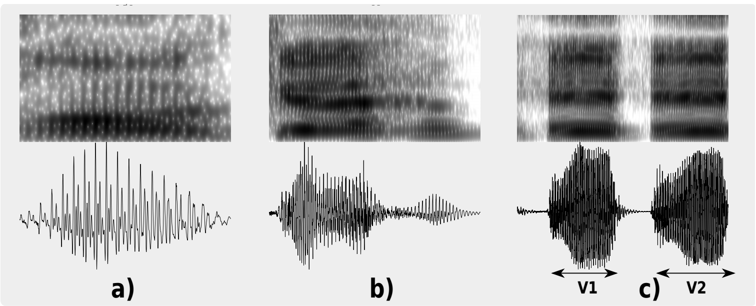

Figure 1: Representation on the spectral (above) and time (below) domains of a) a smiled vowel, b) a chuckling vowel and c) a laughter syllable

Speech-smile is a term used to describe the alteration of speech due to smiling. As already mentioned, smiles can indeed be audibly identifiable in speech [9, 10]. It is therefore possible to make use of this dimension in this work.

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 fricative sound in laughter [13]. In this work, the pattern for laughter syllables can also be described as the succession of two fricatives because this pattern can be found in natural laughs and, in particular, in our data.

Chuckling or shaking vowels, as presented in [12, 7], are vowels altered by some kind of tremolo in an amused sentence. This can be seen by comparing the smiled vowels, chuckling vowels and laughter syllable patterns. Fig. 1 shows the common temporal and spectral representations of a smiled vowel (a), a chuckling vowel (b) and (c) two consecutive laughter syllables found in our data. Some observations can be made out of these figures. First, a discontinuity in the spectral representation of the chuckling vowel can be noticed. A more obvious and accentuated discontinuity can be observed in the laughter syllables spectral representation separating the first vowel and the second one (V1 and V2 respectively, in Fig 1 (c)). Since, in the case of the laugh, this discontinuity is due to the fricative separating the two vowels and considering the fact that both phenomena occur in amused speech (and apparently at two different levels of amusement), the discontinuity in the case of the chuckling vowel must also be due to an air pulse.

After a comparison of the chuckling and smile vowels in the temporal domain, it seems like the chuckling vowel pattern is formed of a sequence of two vowels separated by a ”breathy vowel” (which would be located where the discontinuity is). The conclusion drawn from these observations is that the tremolo-like sound perceived in chuckling vowels is produced by the alteration of a vowel (probably a speech-smile vowel) by a muscular mechanism similar to the one producing laughter (please refer to [13] for more details on laughter production).

2.2. ASC Database

For the purpose of this study, the database used is the same as the one used in [7]. We thus have 1004 laughter syllables, 335 smile vowels and 48 chuckling vowels in total.

The database details can be found in [7]. However for the clarity of this article, Table 1 gives a summary of the datasets from which the ASC were extracted. 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, Ch = chuckling vowels, Laugh = laughter syllables). The recording conditions column contains whether the data recorded are ”clean” or ”noisy” and whether they are naturally expressed or acted.

Dataset	**Lang**	**Rec Cond**	**Sm**	**Ch**	**Laugh**
DS1	French	clean/acted	+	+	-
DS2	English	noisy/natural	+	+	+
DS3	paralinguistic	clean/natural	-	-	+

Table 1: Dataset summary table: In the Lang column, ”paralinguistic” means that there was no specific 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.

2.3. Data Analysis

In this section, some data analysis on the database presented in the previous section will be presented.

First the durations of each amused speech component sample was computed and the density distribution of their values is given in Fig. 2.

As we can see, the durations of laughter syllables tend to be longer than the two other ASCs in our database, and chuckling vowels seem to be on average, longer than the smiled vowels.

In Section 2.1, we pointed out the discontinuity in the spectral representation of the chuckling vowels and of the laughter syllables. In fact this observation

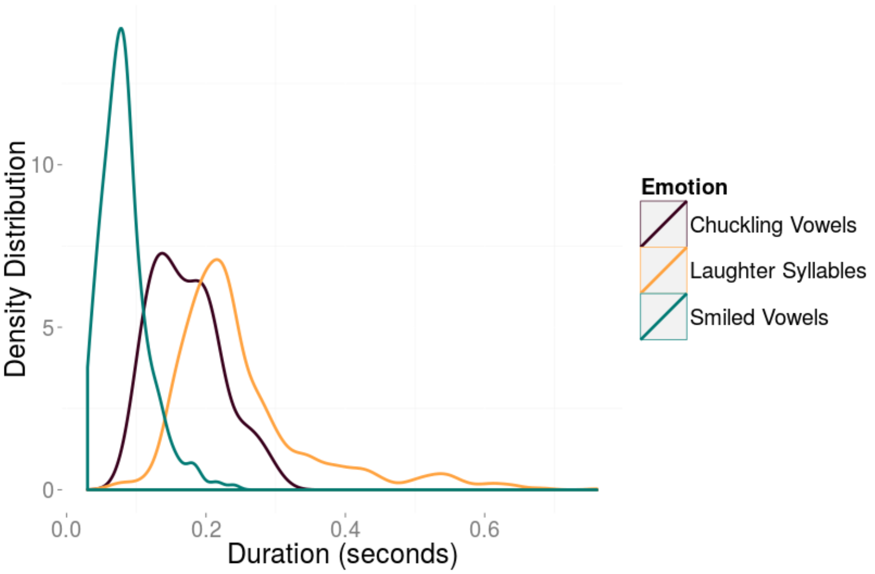

Figure 2: Amused Speech Components durations density distribution.

suggests the speech-smile vowels to have properties such as pitch and energy (or power) that are more stable than the chuckling 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. Indeed, as we know, fricative, breathy and unvoiced sounds in general have low or even null pitch and energy vowels compared to voiced sounds. These two parameters were thus investigated for the three different amused speech components.

Fig. 3 and Fig. 4 show the patterns of the pitch and energy values respectively computed for smiled vowels, chuckling vowels and laughter syllables, plotted over the temporal representations of these components. The temporal representations of the amused speech components showed here, are common patterns that can be found in our database for each of these components. The pitch here was computed using the ESPS method of the Snack library [14] with a window of 20 ms width shifted by 10 ms. The energy was computed using a 10 ms

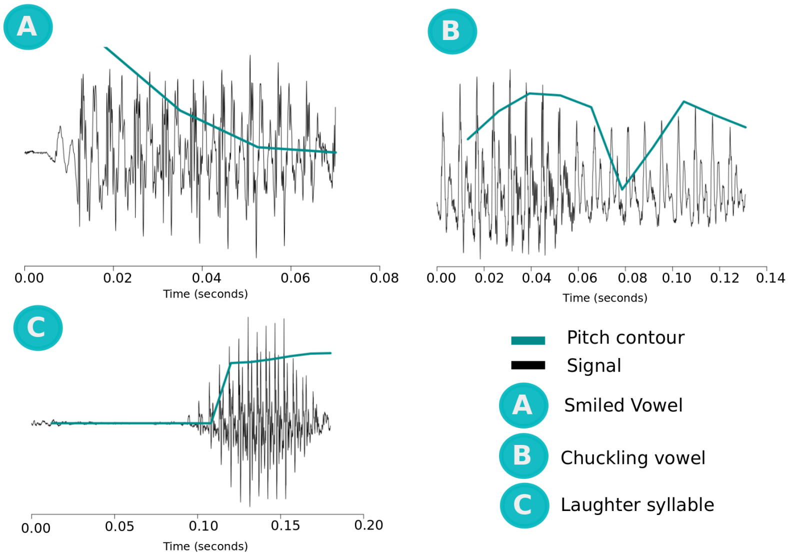

Figure 3: Temporal representations and pitch contours examples for different amused speech components

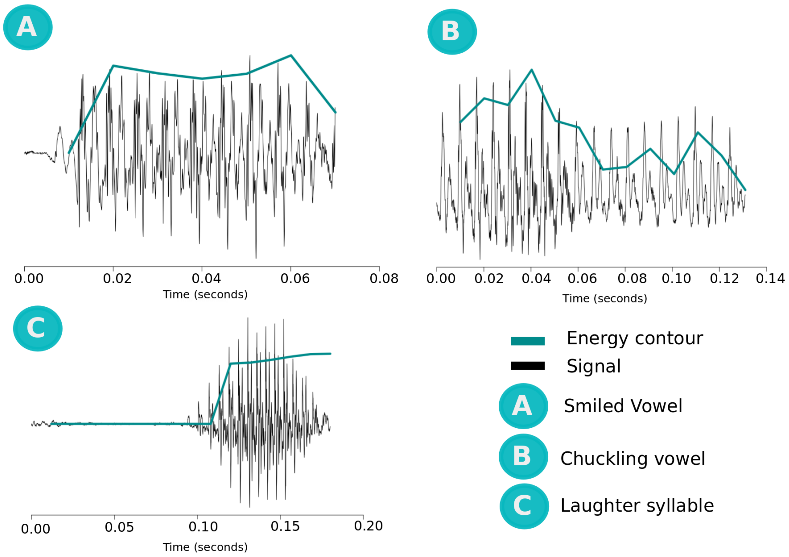

Figure 4: Temporal representations and energy contours examples for different amused speech components

windows shifted by 10 ms. In these plots, the actual values of the pitch and energy are not represented because their purpose is to focus on the pitch and energy patterns with respect to the temporal waveforms.

As can be seen, and as expected, the pitch and energy tend to be rather monotonous for smiled vowels compared to more variant shapes for chuckling vowels and an almost binary shape for laughter syllables, passing from a low value during the fricative part of the laughter syllable to a high value during the voiced part of it. These features will be therefore, the basis of the features used for our systems later on.

3. Stability-Bases Features Efficiency

In our previous work [12, 7], different sets of features were studied for the classification of smiling and chuckling vowels, and for the classification of all the ASCs defined here respectively. In these studies, we introduced Stability-based Features (SF) under the hypothesis that some chuckling vowels characteristics (such as the pitch and the signal power/energy) might be less stable than the smiled vowels ones. The SF proved to give good discriminative results. They also proved to be discriminative for the previously mentioned ASCs and laughter syllables.

3.1. Stability-Based Features

The Stability-based Features were inspired from the observations made in Section 2.1 and Section 2.3.They were originally based on the stability of the pitch and signal power values.

So, in order to represent this stability, the pitch is first estimated on each sample of smile vowel, chuckling vowel and laughter syllable using the ESPS method of the Snack library [14]. This is done, as above, 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 (see Section 4). 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 temporal log-power of the signal is computed using the following formula:

$$P(i) = 10 \log \frac{x(i)^2}{\Delta T} \qquad (1)$$

$x(i)$ being the i^{th} signal sample and ΔT 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. Previous Experiment

3.2.1. Experiment Description

In the former experiment [7], a combination of the SF and commonly used features in speech and emotion recognition were used to train different systems (kNN, SVM, and neural network) to classify the three ASC types. These commonly used features are based on the MFCCs and on the pitch estimation. The pipeline shown in Fig. 5 was applied to each of the system compared. The same pipeline will be used further in this work too.

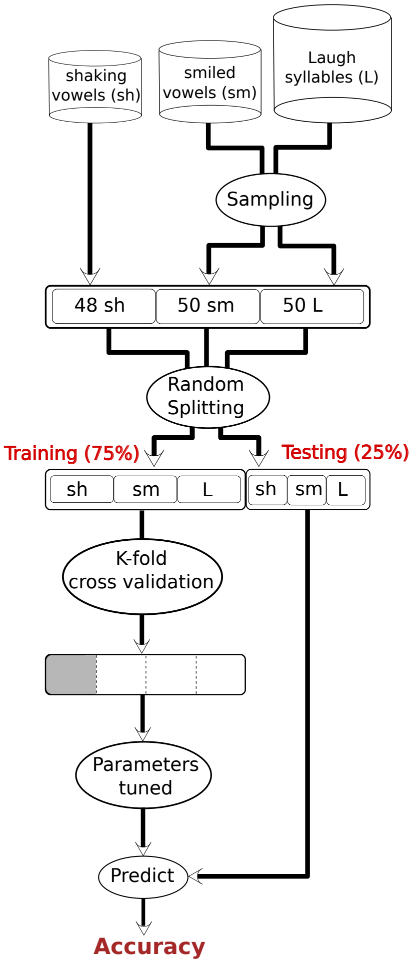

Figure 5: Data splitting, systems training and systems testing pipeline.

To tackle the imbalanced classes problem (48 shaking vowels 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 the accuracy is computed.

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

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.

Each system's hyperparameters values/dimensions varied for every iteration of the process described in Fig. 5. So, several values of the corresponding parameters for each system were used during the tuning step, and only the ones with the best performance was kept for retraining the system later on.

3.2.2. Classification Efficiency

From the previous work, we first concluded that using the SF alone gave better results than using the MFCC or the f0 vectors alone with any of the systems presented. When combining the features, the results showed 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 significantly better results than all other systems and the single layer neural network's results were significantly better than any other system when using non-combined feature vectors.

When comparing the systems, the SVM with polynomial kernel obtained the best score when used with the SF+f0 combination (78.96%). Considering only one feature vector (higher table) the best score was obtained by the neural network when used with the SF feature vector (76.2%).

The results obtained using the SF for this classification task encouraged us to continue investigating features computed from the pitch and energy in order to improve our system.

4. Improved Feature Sets

As expected, the SF features based on the varying stability of the pitch and power for the three amused speech components, successfully classified these latter significantly better than chance. These results, encouraged to improve the SF feature set.

4.1. Improved Stability-Based Features Set

In the following, the temporal log-power described in equation 1, was replaced with the simpler temporal log-energy since it showed to give the sames results because the temporal energy is equal to the temporal power within a constant, and its simplicity makes it more computationally efficient. The temporal log-energy equation is given by:

$$E(i) = 10 \log x(i)^2 \tag{2}$$

Thus the power-based SF features are replaced by the energy-based SF features by only replacing equation 1 by 2 .

After this first modification, improving the SF feature set begins by considering the mean value of the residuals extracted from the pitch and energy. These two new features are added to the previously described SF. Fig. 6 shows the density distributions of the mean and standard deviation values of the pitch and energy log value per amused speech component.

From these plots, we can see that the mean of the pitch residuals seems actually to be the most discriminating feature among them since the distributions for this feature are the less overlapping. The log energy residuals standard deviation values seem to be discriminative between the speech-smile syllables and the other two classes while the pitch residuals standard deviation and the mean log energy residual features tend to be more discriminative between the laughter syllables and the other two classes.

Secondly we also consider other pitch-based features: the jitter and shimmer features. These features represent the cycle-to-cycle fundamental frequency and

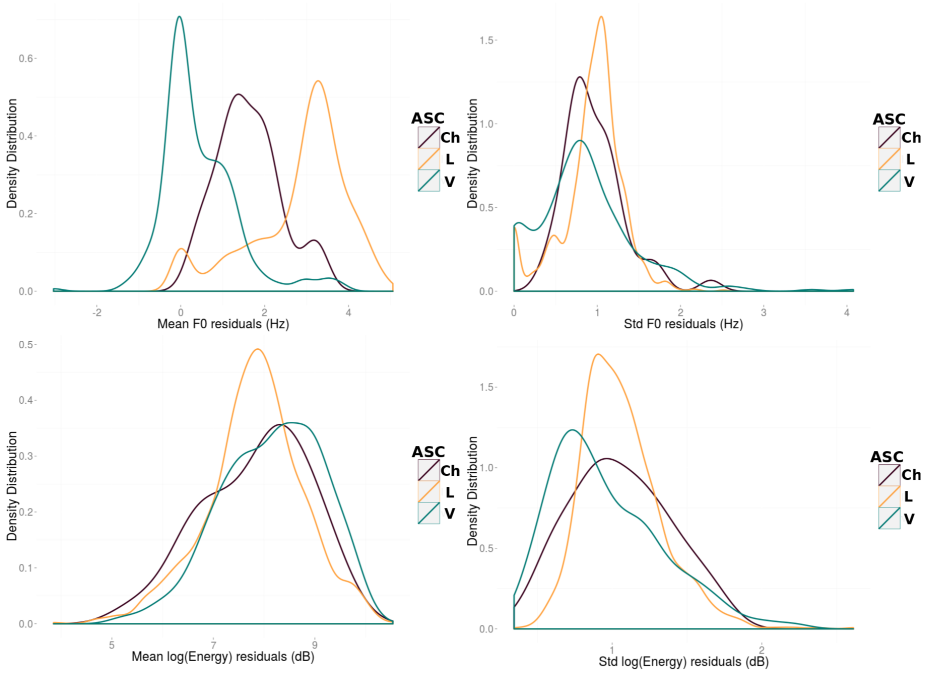

Figure 6: Density distributions of the mean and standard deviation values for the pitch and the log value of the energy per Amused Speech Component (ASC).

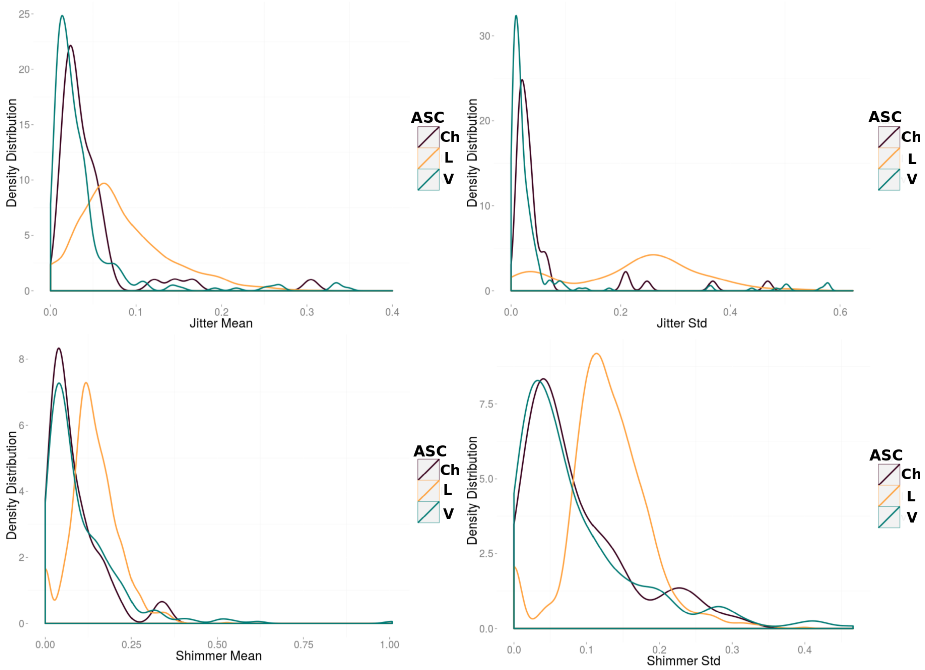

Figure 7: Density distributions of the mean and standard deviation values for the jitter and shimmer for each Amused Speech Component (ASC).

amplitude variations respectively [15]. They proved to be efficient in audio classification of emotions [16].

These were computed using the description given in [16]. Their calculation was made using the pitch and peak normalized cross-correlation value extracted using the Snack library. Again, here, a 20 ms wide window was used shifted by 10 ms in order to compute them. Fig. 7 shows the distribution of the mean and standard variation calculated for each of the extracted jitter and shimmer. The distributions are plotted per amused speech component for each of these statistics.

After observing these graphs, we can see that these features could potentially contribute at discriminating mostly between the laughter syllables and the other two classes. So, these 4 features obtained are also added to the SF aforementioned.

The new SF will be referred to as **newSF** in the remaining of this article.

4.2. Statistical Feature Set

Also based on the pitch and the energy, another feature set was extracted using the OpenSMILE feature extraction tool [17]. Using this tool, different features are extracted from the energy, the pitch, the jitter and the shimmer.

First the contours of the root mean square and log energy are extracted from the amused speech components with a 10 ms window shifted bu 10 ms. The pitch contour is estimated using the sub-harmonic sampling method with a 40 ms wide window shifted by 10 ms. From this contour, the jitter and shimmer contours are also computed. All these contours are then smoothed.

A large set of features is then extracted from each of the 2 energies, the pitch, the jitter and the shimmer smoothed contours and each amused speech component:

- the range
- the position of the maximum and minimum values of each signal
- the arithmetic mean
- all the features listed in [17] concerning the quadratic and linear approximation of the segment contours.
- statistical moments features, more specifically the standard deviation, the kurtosis and the skewness values.
- all the time statistics in frame values, such as "the time during which the signal is above or below a certain percentage of its total range"

A total of 810 features were extracted for each amused speech component. This set of feature will be referred to as **OSfeats** in the following.

5. Improving the Classification System

The two new feature sets presented in the previous section are used for the classification task of the amused speech components. They are used with the NN and the SVM with polynomial kernel since these two methods gave the best results in our previous classification attempt. To these will be added an SVM with Radial Basis Function (RBF) kernel (the features with the best results obtained in our previous attempt will be retrained with this algorithm too for the sake of comparison). The pipeline for data splitting, training and testing steps is the same as the one described in Section 3.2.1 and in Fig. 5.

5.1. Principal Component Analysis

Concerning the OSfeats, first results were obtained training each of the three aforementioned methods with the whole 810 OSfeats features. Even though the average accuracy values obtained were better than chance, they were lower than the previous results obtained (58,2% when the NN is used, 60.1% and 60.04 when using an SVM with polynomial and RBF kernels respectively). This is due to the systems overfitting the data because of the high number of features dimensions with respect to the systems used. So, a Principal Component Analysis (PCA) was applied on this set of features for dimensionality reduction. The resulting set of features was used for training and testing, and this, for each of the NN and SVMs algorithms. The same pipeline as in Fig. 5 was applied here too, but 100 repetitions each time instead of 1000.

Fig. 8 shows the average accuracy obtained per model and per number of principal components used. On the left column of this figure, the mean accuracy values are shown. They are obtained after training a neural network (A), an SVM with a Radial Basis Function kernel (B) and an SVM with a polynomial kernel (C) on the features after their dimensions were reduced to N via PCA. They are plotted with respect to the number of principal components N. A line was then fitted to the obtained mean accuracy values using a third order polynomial equation. In order to estimate the optimum dimension to use, the derivative of the fitted lines is computed and plotted on the right side of the figure.

For each case A, B or C, we chose the dimension corresponding to the first absolute accuracy difference lower than 0.02%. Indeed, for each curve, this would correspond to the point on the fitted curve before this latter stabilizes (corresponding in our case to an absolute difference value lower than 0.02%) and then decreases (in the case of B and C). This chosen value for each case also has a high enough accuracy value for each of the three cases. These values were 58 components for the neural network, 70 for the SVM with polynomial kernel and 62 for the SVM with RBF kernel. It is worth to be noted though, that the systems generally have good performances even with a very small number of principal components used for the dimensionality reduction.

5.2. Improved Results

Table 2 contains the average accuracy values after training each set of features with a feedforward neural network (NN), an SVM with RBF kernel and an SVM with polynomial kernel.

System	**SF**	**SF+f0**	**newSF**	**OSfeats**
NN	76.2%	78.1%	80.18%	81.4%
SVM-RBF	74.2%	79.1%	78.4%	82.3%
SVM-Poly	75.6%	78.96%	79.3%	**82.8%**

Table 2: Mean accuracy results of each system per feature set. The best result is a statistically significant result under a 95% CI Student's t-test (in bold).

The first point to be noted from this table is that the OSfeats feature set outperforms the other sets of features. The best obtained result was the one obtained with the SVM-Poly trained with the OSfeats. In order to check its statistical significance a Student's t-test was applied on the accuracy distribution obtained with this system. The test compared the distribution to all the distributions obtained with all other systems in Table 2. The statistical significance of this result could thus be verified with all other systems except with the

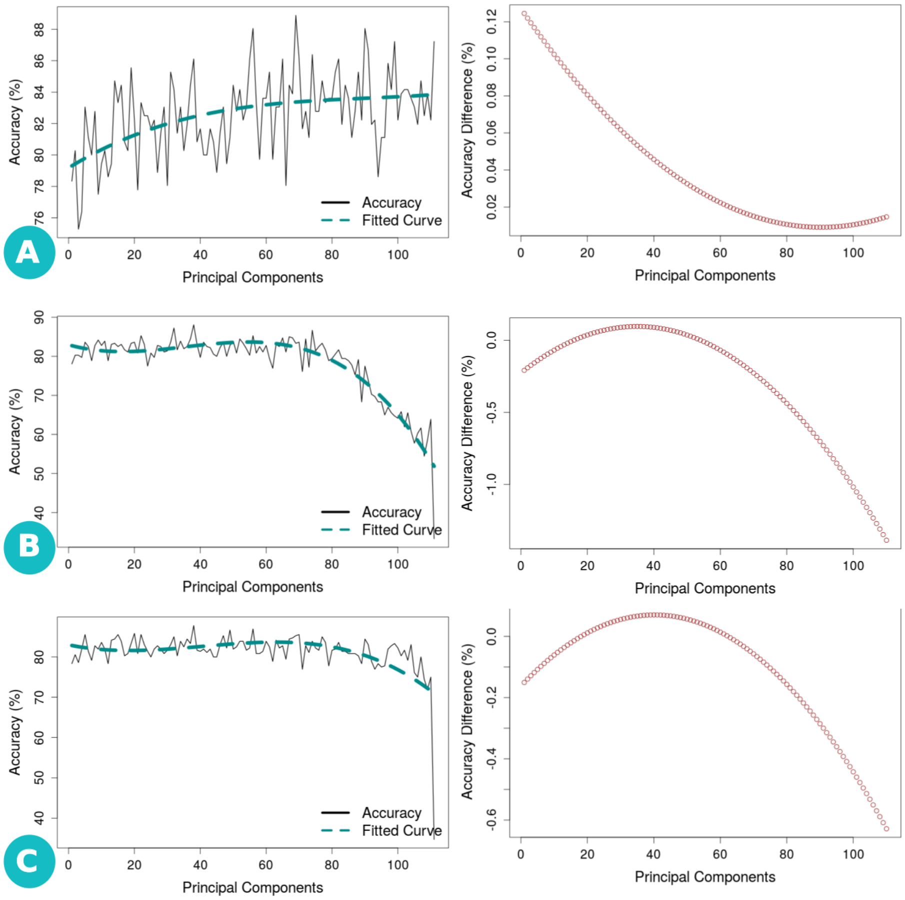

Figure 8: Mean accuracy values obtained with a neural network (A), an SVM with a Radial Basis Function kernel (B) and an SVM with a polynomial kernel (C) after dimensionality reductions are applied to the features.

OSfeats trained with the SVM-RBF (for which we obtained a p-value of 0.12).

The second point to be noted is that the newSF features generally outperform SF and SF+f0 features (except for the SVM-RBF trained with SF+f0, which outperforms SVM-RBF trained with newSF).

It is also worth noticing that the best accuracy value obtained by training a NN with the newSF features outperforms all accuracy values obtained with the SF and SF+f0 features.

Since this is a multiclass classification problem, in order to analyze these results more in-depth, Table 3 shows the Area Under the Receiver Operation Characteristics (AUROC) curve for each of the classes (chuckling vowels, smiled vowels and laughter syllables), per system and per feature set. Calculating the AUROC was made by dealing with this multiclass problem as a one-vs-all problem.

Class	**System**	**SF**	**SF+f0**	**newSF**	**OSfeats**
Ch	**NN**	0.794	0.803	0.82	0.891
	SVM-RBF	0.758	0.875	0.853	0.871
	SVM-Poly	0.754	0.798	0.888	**0.898**
Sm	**NN**	0.792	0.819	0.83	0.891
	SVM-RBF	0.76	0.844	0.879	**0.904**
	SVM-Poly	0.747	0.854	0.866	0.872
L	**NN**	0.885	0.871	0.895	0.923
	SVM-RBF	0.892	0.869	0.899	0.899
	SVM-Poly	0.871	0.872	0.921	**0.934**

Table 3: AUROC per class with bold being highest value per class. Ch = chuckling vowels, Sm = Smiling vowels and L = Laughter syllables

From this table, we first notice that all the results are higher than 0.75 and that 83% for the values are higher than 0.8 which shows that all these feature sets work well for each of the classes. We also notice that the newSF and the OSfeats sets are the only ones showing AUROC values higher than 0.8 in all cases. The second point we can note is that the results are generally similar to the accuracy results of table 2: the OSfeats set always show the best results and the best results obtained from the newSF set is better than all results from the SF and SF+f0 sets for each class.

In general, we can see that the feature sets discriminate the laughter syllables class (L) better than the other classes (Ch and Sm). Indeed this is coherent with the features discrimination efficiency of the laughter syllables compared to the others, visible in figures 6 and 7.

From all these results it is safe to conclude that in this study the OSfeats give the best results followed by the newSF feature set. But, although the former outperforms the latter, the newSF feature set has the advantage of being interpretable and not requiring a PCA to be applied, which is computationally more efficient. More effort should be put to discriminate between the smiling and chuckling vowels. Doing this will most probably increase the efficiency of the whole system for this task.

6. Conclusion and Perspective

In this paper, we first introduced Amused Speech Components which will be used in future work to assess amusement level in one's speech. A summary of our previous work regarding classifying these ASC was then given. This has the ultimate goal of amusement intensity level assessment. We also push the analysis of the data previously gathered further than in our previous work in order to have a better understanding of these ASC. In this work we improve our previous classification results by creating new sets of features.

One limitation to be noted is that the majority of the features presented here, discriminated individually better the laughter syllables from the other classes than the chuckling or smiled vowels from the other classes. Another limitation is the amount of chuckling vowels data available. Indeed, this prevents us from using systems that have proved to be very efficient for classification tasks but require a larger amount of data.

We plan on tackling these limitations by first gathering a larger database.

Then we will attempt to use systems that deal efficiently with time dependent data, hopping to get better results (systems such as Hidden Markov Models, or Recurrent Neural Networks and more specifically Long Short-term Memory networks).

Finally we plan on building an ASC detection system. Among others, the detected ASC would be used to assess a subject's amusement intensity level.

Acknowledgment

The authors would like to thank Dr. Payton Lin for the constructive discussions regarding the topic of this article.
This work was partly supported by the Chist-Era project JOKER with contribution from the Belgian Fonds de la Recherche Scientifique (FNRS), contract no. R.50.01.14.F.
H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

References

[1] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, in: INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, 2014, pp. 223–227.

[2] E. Yuncu, H. Hacihabiboglu, C. Bozsahin, Automatic speech emotion recognition using auditory models with binary decision tree and svm, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, 2014, pp. 773–778.

[3] T. Pfister, P. Robinson, Speech emotion classification and public speaking skill assessment, in: A. Salah, T. Gevers, N. Sebe, A. Vinciarelli (Eds.), Human Behavior Understanding, Vol. 6219 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 151–162.

[4] A. S. Patwardhan, G. M. Knapp, Affect intensity estimation using multiple modalities, CoRR abs/1607.01075.

[5] Y. Arimoto, S. Ohno, H. Ida, An estimation method of degree of speaker's anger emotion with acoustic and linguistic features, Journal of Natural Language Processing 14 (3) (2007) 147–163.

[6] A. Dhall, R. Goecke, Group expression intensity estimation in videos via gaussian processes, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 3525–3528.

[7] K. El Haddad, H. Cakmak, S. Dupont, T. Dutoit, Towards a Level Assessment System of Amusement in Speech Signals: Amused Speech Components Classification, in: IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Abu Dhabi, UAE, 2015.

[8] K. El Haddad, H. Cakmak, S. Dupont, , T. Dutoit, Towards a Speech Synthesis System with controllable Amusement levels, in: 4th Interdisciplinary Workshop on Laughter and Other Non-Verbal Vocalisations in Speech, Enschede, Netherlands, 2015, pp. 15–18.

[9] A. Drahota, A. Costall, V. Reddy, The vocal communication of different kinds of smile, Speech Commun. 50 (4) (2008) 278–287.

[10] V. Tartter, Happy talk: Perceptual and acoustic effects of smiling on speech, Perception & Psychophysics 27 (1) (1980) 24–27.

[11] J. Trouvain, Phonetic aspects of "speech laughs", in: Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L'Harmattan, 2001, pp. 634–639.

[12] K. El Haddad, S. Dupont, H. Cakmak, T. Dutoit, Shaking and Speech-smile Vowels Classification: An Attempt at Amusement Arousal Estimation from Speech Signals, in: IEEE Global Conference on Signal and Information Processing (GlobalSIP), Orlando, Florida, US, 2015.

[13] W. Ruch, P. Ekman, The expressive pattern of laughter, in: A. W. Kaszniak (Ed.), Emotion, qualia, and consciousness, Word Scientific Publisher, Tokyo, 2001, pp. 426–443.

[14] K. Sjölander, The Snack Sound Toolkit [computer program webpage] (consulted on September, 2014).

[15] M. Farrus, J. Hernando, Using jitter and shimmer in speaker verification, IET Signal Processing 3 (4) (2009) 247–257.

[16] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, J. D. Newman, Stress and emotion classification using jitter and shimmer features, in: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Vol. 4, 2007, pp. IV–1081–IV–1084.

[17] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM International Conference on Multimedia, MM '13, ACM, New York, NY, USA, 2013, pp. 835–838.

Authors Biography

Kevin El Haddad obtained a Masters degree in Microsystems and Embedded systems (with honors). He is currently pursuing a PhD in the TCTS Lab of the University of Mons (Belgium) under the supervision of Prof. Thierry Dutoit and Dr. Stéphane Dupont. His research interests include multimodal synthesis and recognition of affect signals, affective computing and machine learning applications for Human-Agent Interactions.

Hüseyin Çakmak holds a double degree in Aeronautics from the Higher Institute of Aeronautics and Space (ISAE) and in Electrical Engineering from the Polytechnic Faculty of Mons (FPMS). He obtained his PhD in 2016 under a FRIA grant. He is now a post-doctoral researcher at the TCTS Lab of UMONS. His research interests are in machine learning and affective computing.

Stéphane Dupont PhD degree in EE in 2000. Post-doctoral associate at ICSI (California) in 2001-2002, working towards robust speech recognition ETSI standard. Head of ASR research at Multitel (Belgium) in 2002-2008. Joined University of Mons in 2008. Contributing to several industrial, regional and EU projects. Head of research group on machine intelligence. Holds 3 international patents and co-authored 120 papers.

Thierry Dutoit is a full professor at UMONS, Belgium. Between 1996 and 1998, he spent 16 months at AT&T-Bell Labs, in Murray Hill (NJ) and Florham Park (NJ). He initiated the MBROLA project for free multilingual speech synthesis, the eNTERFACE workshops, of which he coordinates the steering committee, and UMONS/NUMEDIART Institute for Creative Technology, of which he is the director.