July 20, 2016

Report

Table of content

1. Laugh Event
2. Acoustic system visualisation
3. Performance

1. Laugh events

Estonian dataset	DigiSami conversation dataset

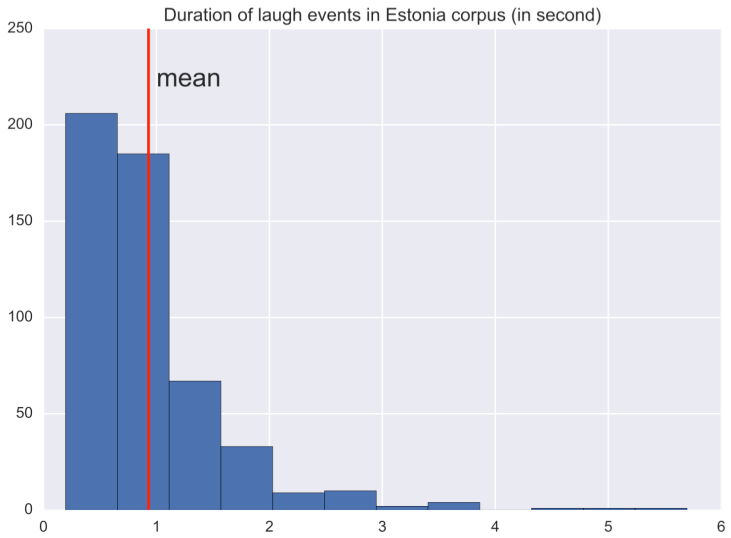

Fig 1a

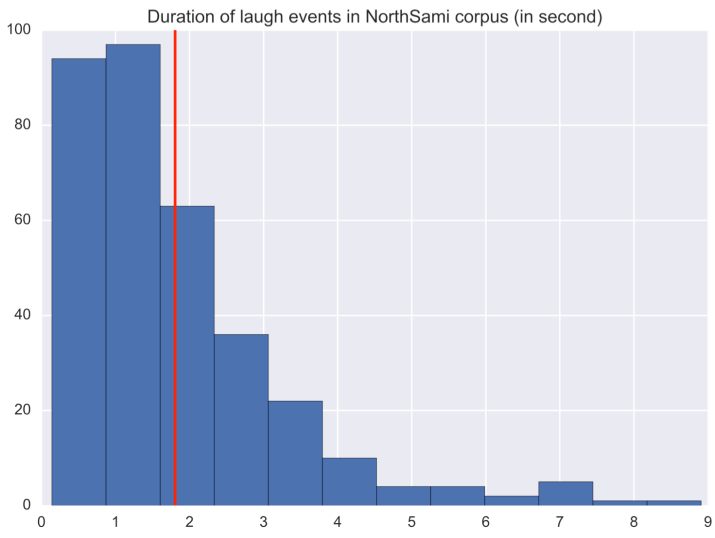

Fig 1b

From the duration point of view, Fig.1a shows the histogram of laugh duration in Estonia dataset. Most of people laughed for approximated 0.8s, and the laughing is rarely longer than 2 second.

Conversely, Fig. 1b indicates people in DigiSami conversation often laugh for longer duration (i.e ~ 1.8s). Most of the events are located from 0.1 to 2 seconds and longer than 4 second laugh is rarely happen.

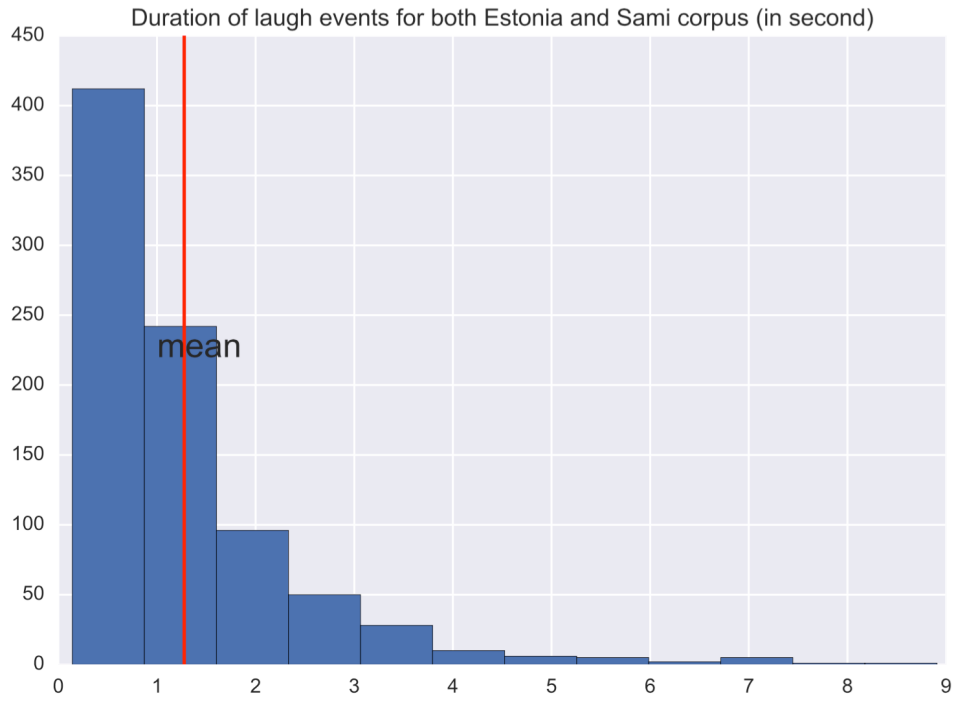

Fig 1c

In general, most of people laugh from 0.1 to 1.6 second with the mean value of 1.2 second, and there are tremendous amount of short laugh events during the conversation. There also exist significant different between laughing events in Estonia dataset and DigiSami dataset, which is the results of following factors:

- The scenarios of experiment setup (* can you help me give some detail here)
- The familiarity between people in the conversation
- The culture difference between Estonia and North Sami ? (is it possible ?)

Estonian dataset

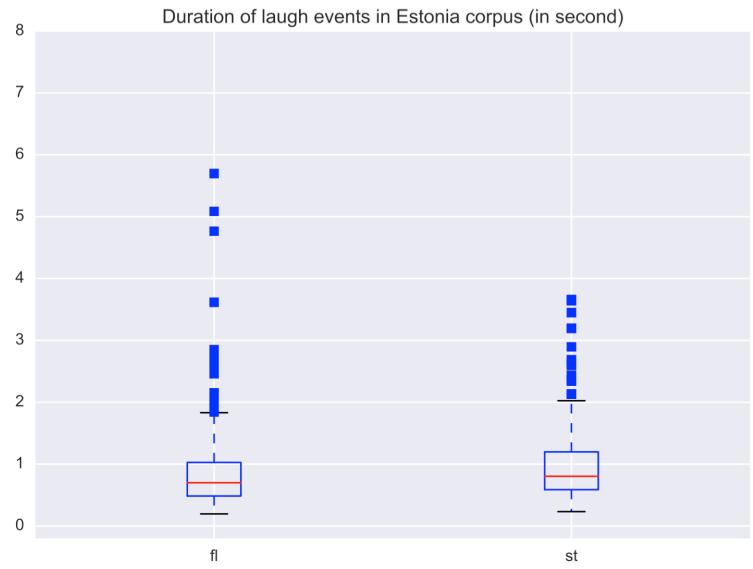

Fig 2a

Fig. 2a illustrates the differences of the distribution between free-laughter (fl) and speech-laugh (st) in Estonian dataset. We can see that speech-laugh is slightly longer than free-laughter, and is frequently appeared during the conversation than free laugh event

DigiSami conversation dataset

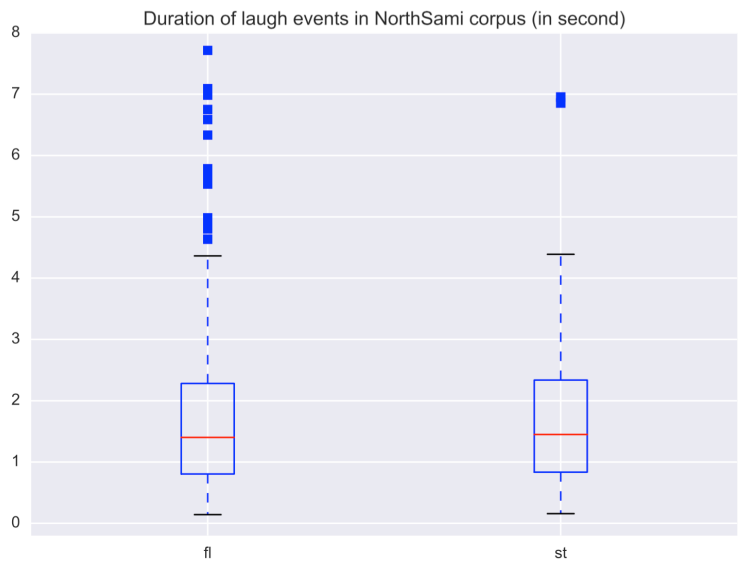

Fig 2b

However, both events are equally distributed in the NorthSami corpus. Moreover, laugh events in NorthSami is widely distributed than Estonia data, but the number of outlier in NorthSami is smaller than in Estonian corpus.

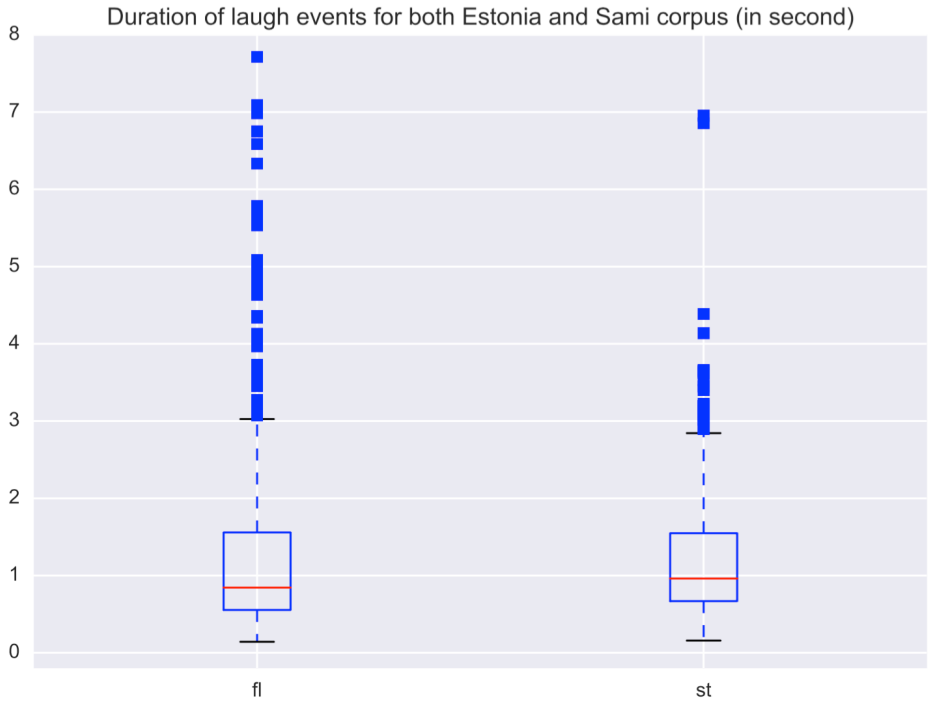

Fig 2c

Taking into account both events from Estonia and DigiSami, we can see that free-laughter are widely distributed than speech-laugh (i.e. longer duration range), but the average duration of fl is 0.1 second shorter than st. On the other hand, the number of outlier in free-laughter is greater than speech-laugh which emphasises the unpredictable of free-laughter events.

Estonian dataset

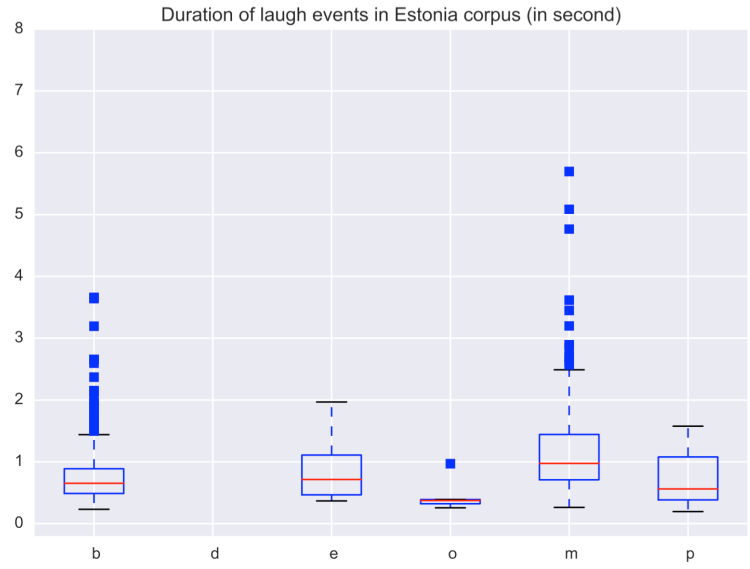

Fig 3a

There are 7 emotional states that affects the laugh events:

- b: breath heavy breathing, smirk, sniff;
- e: embarrassed speaker is embarrassed, confused,
- m: mirth fun, humorous, real laughter,
- d: derision mocking the partner
- p: polite polite laughter showing positive
- o: other laughter that doesn't fit in the

In Estonia corpus, most of the events is humorous laugh, polite laugh, embarrassed laugh and sniff laugh. However, the humorous laugh and smirk laugh is often unexpectedly longer with more outlier.

DigiSami conversation dataset

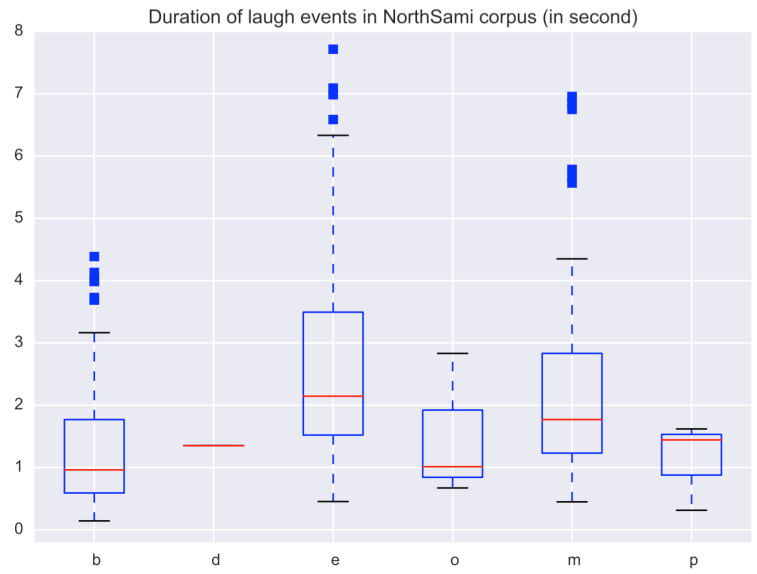

Fig 3b

In NorthSami, we can see much more emotional states in the laugh, and there are an interesting amount of embarrassed laugh during the conversation which can be explained by low acquainted level of the participants.

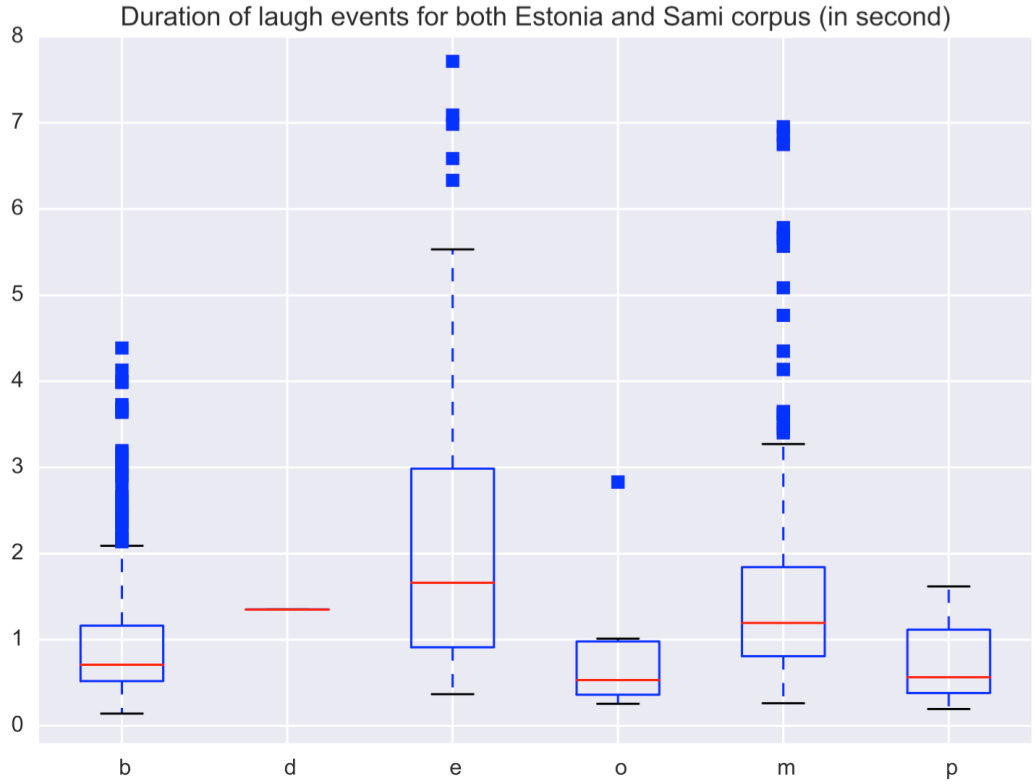

Fig 3c

In general, the participants rarely laughed for mocking their opponents, and since many conversation are between strangers, the embarrassed laugh are often happened (not sure if this is right).

Estonian dataset	DigiSami conversation dataset

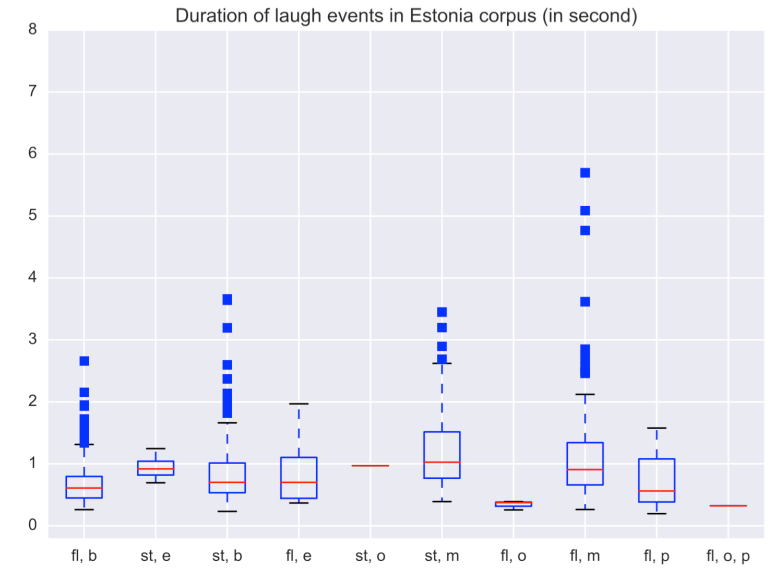

Fig 4a

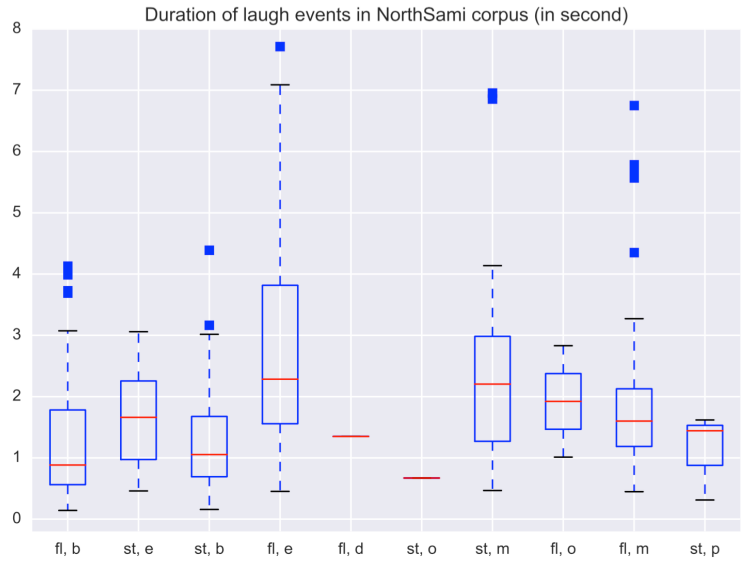

Fig 4b

In Estonian corpus, the most common events is humorous laugh which includes both free-laugh and speech-laugh, conversely, embarrassed free-laugh is the most popular one in NorthSami corpus. Laugh is generally longer in Sami dataset than Estonia dataset. For both corpus, embarrassed laugh is often free-laugh, and humorous laugh is frequently longer with speech-laugh events.

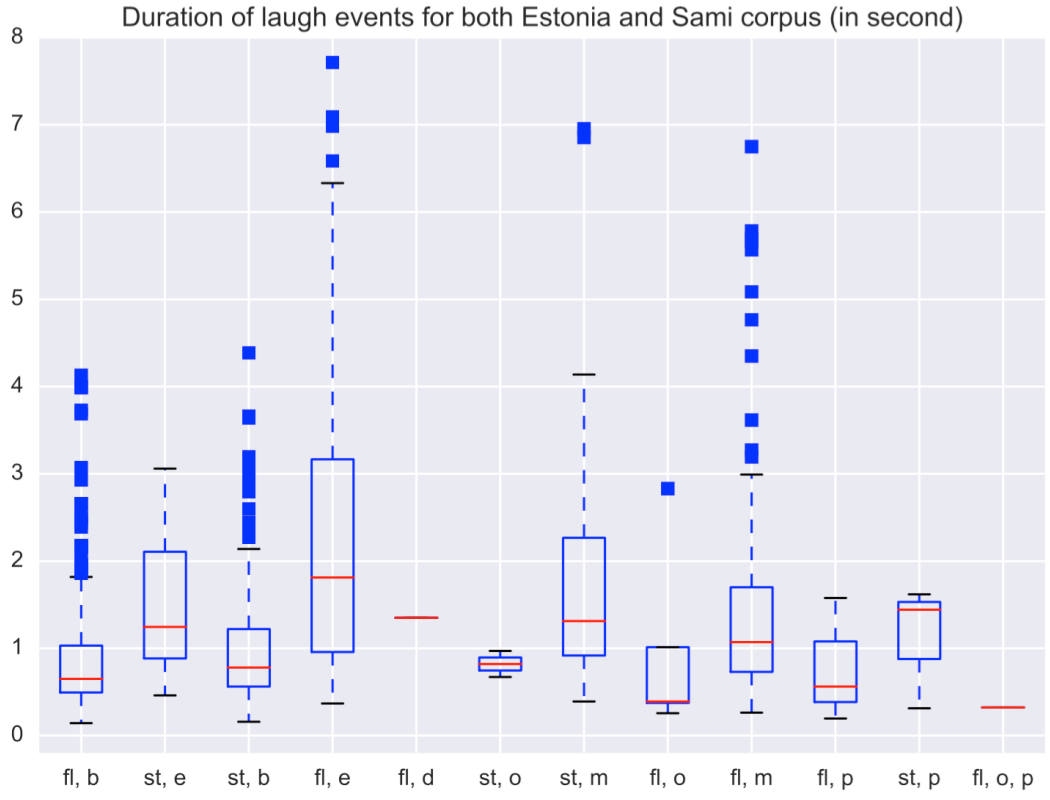

Fig 4c

Combining both dataset into Fig. 4c, we can see free-laugh is distributed in longer range than speech-laugh. However, there is contradictory between embarrassed laugh and humorous laugh. A speech laugh which is results of humorous action often last longer than free-laugh with the same condition, and an embarrassed laugh is last longest with free-laugh speakers.

2. Acoustic system visualisation

Note that the acoustic features is highly non-linear, contradictory, LDA and PCA is linear dimension reduction method. Hence, the new projected space probably cannot capture all the discriminative properties between laugh and non-laugh signal.

Fig. 4 and Fig. 5 illustrate the important of context window length in discriminate between laugh and non-laugh signals, as well as the differences between NorthSami and Estonian corpus. Each MFCC features are processed using a window of 25ms on input audio, and we shift the window every 10ms for the next features. We can see that there are significant amount of laugh events which are longer than 0.25ms (Fig. 1), hence, 1 window of MFCC might not enough to capture all necessary information that characterises the laughing. As a results, we group multiple windows and stack them into 1 big features, the number of surrounding windows are called “context windows”. For a context length equal to 10, it means we use 5 context windows in the past, and 5 context windows in the future to create a “super vector” features. Fig. 4 and 5 highlights the roles of context in laugh signal recognition. We can see that the longer the context the further the non-laugh (red) and laugh (blue) events are pushed into 2 sides of the figure, which is especially applied for Estonian dataset. On the other hand, the acoustic features of laugh signal in North Sami is more difficult to be separated from non-laugh signal compared to Estonian.

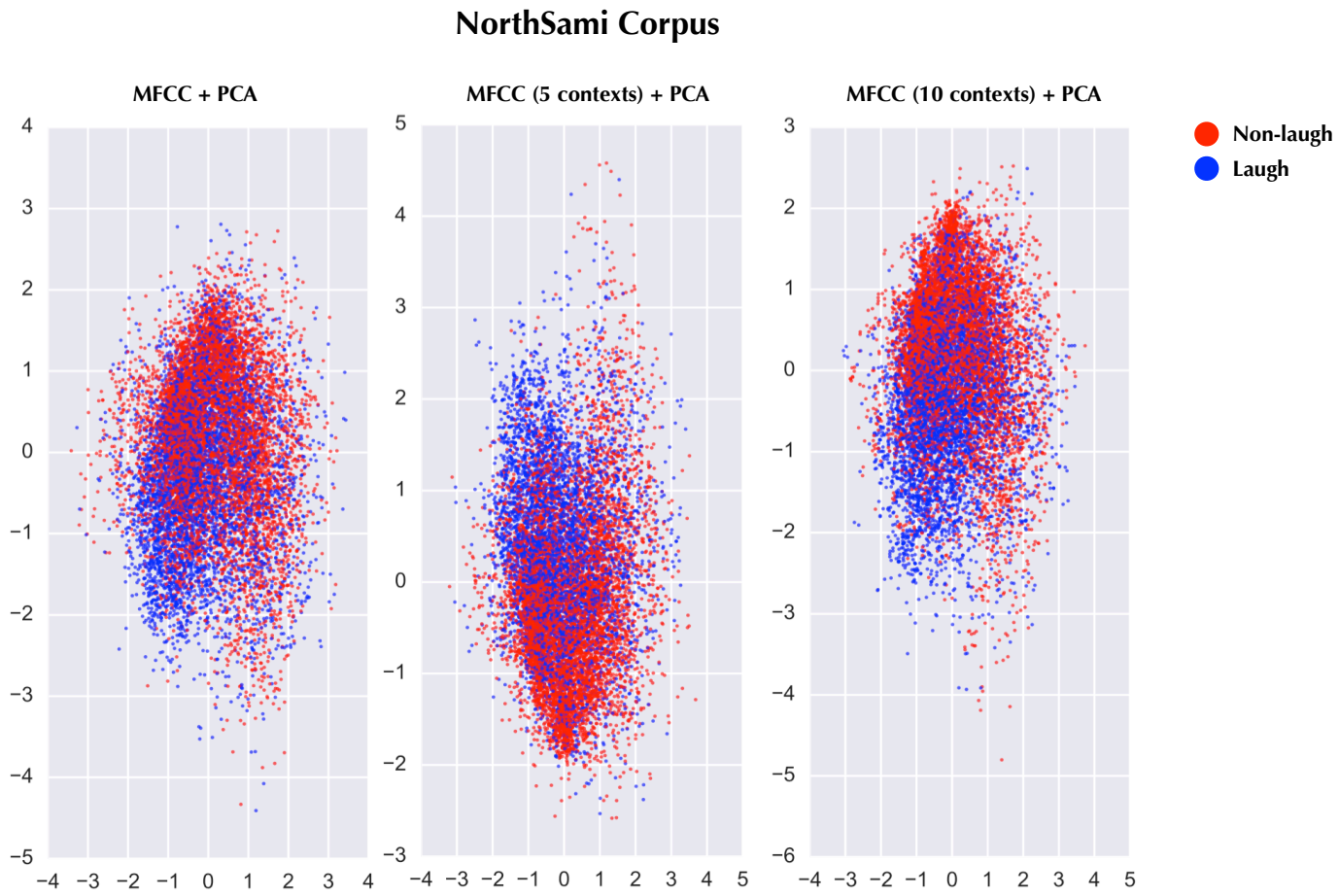

Figure 4: Applying Principal Components Analysis (PCA) on MFCC features with different amount of context windows for NorthSami Corpus.

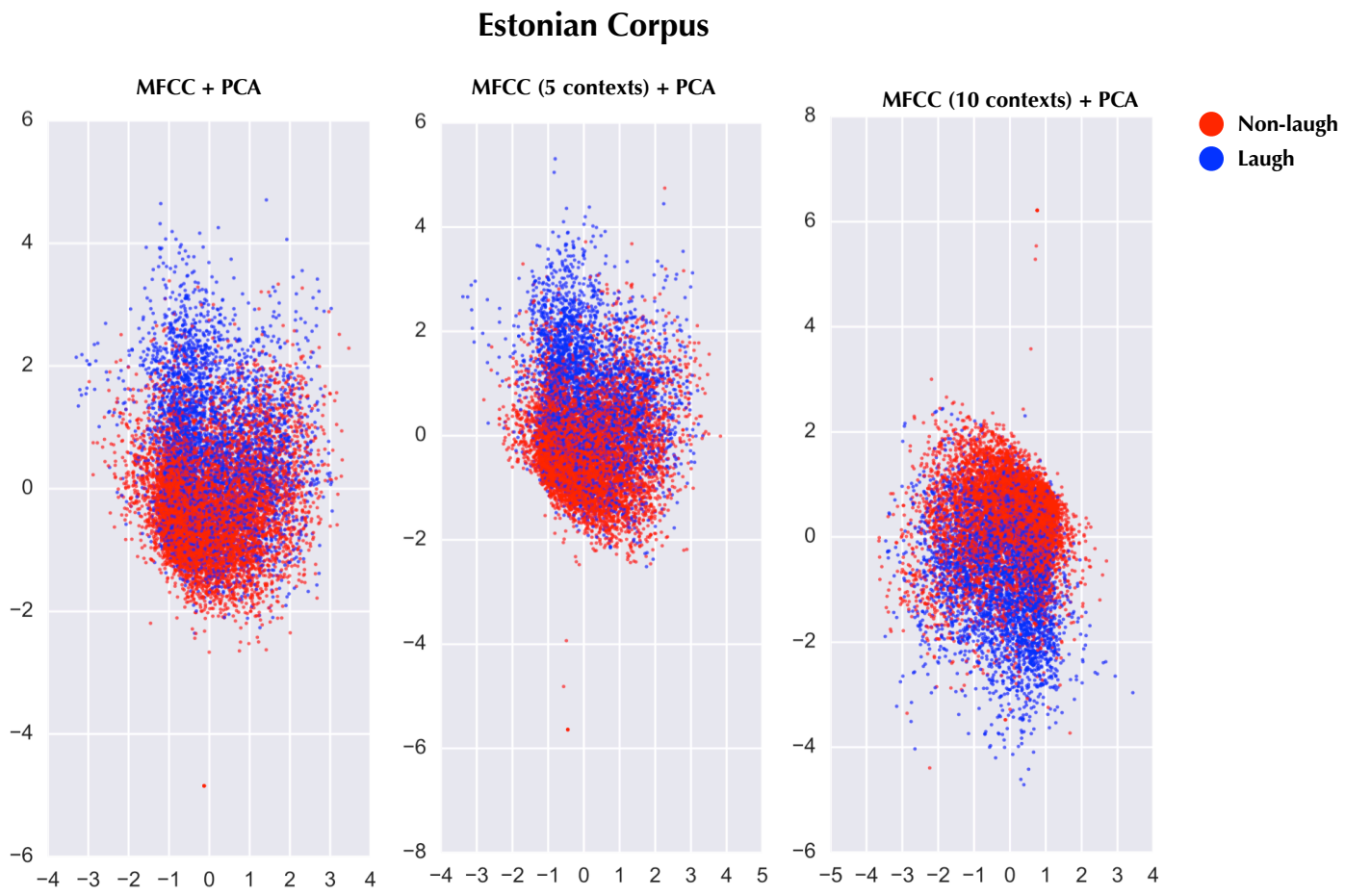

Figure 5: Applying Principal Components Analysis (PCA) on MFCC features with different amount of context windows for Estonian Corpus.

Fig. 6 illustrates the effect of 2 algorithms on different features. We can see that non-laugh and laugh signal are more clearly separated in the case of MFCC with PCA, using pitch features introduces more confusion between the 2 types of signals. However, both of them are clearly separated using LDA, hence, MFCC and pitch can probably used for training a classification between laugh and non-laugh events.

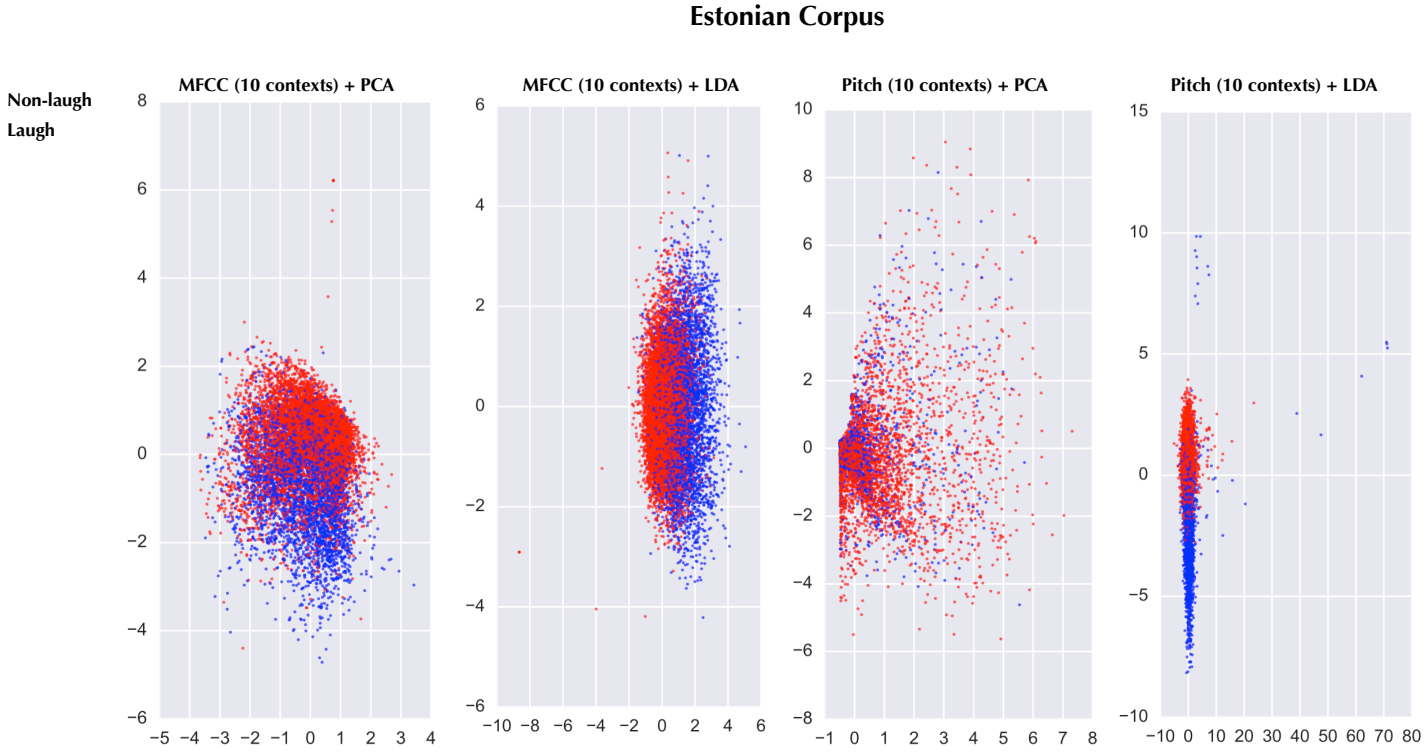

Figure 6: Applying PCA and LDA on MFCC and PITCH as features

Fig. 4, 5 and 6 emphasises the important of features in detecting laugh event from audio signal. We further investigate the influence these features on different types of laugh and dataset which is illustrated in Fig. 7 and 8. In Fig. 7, speech-laugh is clearly separated from free-laugh using LDA for 13 different laugh annotation (i.e. the labels are mixed of laugh types and emotional states). As a results, we can see that the laugh types are seemingly inferred given the mixed information of laugh and emotional states.

However, Fig. 8 indicates the difficulty in extract emotional states information from all 13 annotations. We highlight the dense area of the three most popular states: breath, embarrassed and mirth. The circles are overlapped which illustrates strong confusion between different emotion.

On the other hand, we can see the green zones of Fig. 7 match the mirth zones of Fig. 8, hence, there exist strong relationship between speech-laugh events and mirth emotional state. Conversely, free-laugh is mixed of breath and embarrassed emotion, which make it more overlapped with non-laugh and speech-laugh signal.

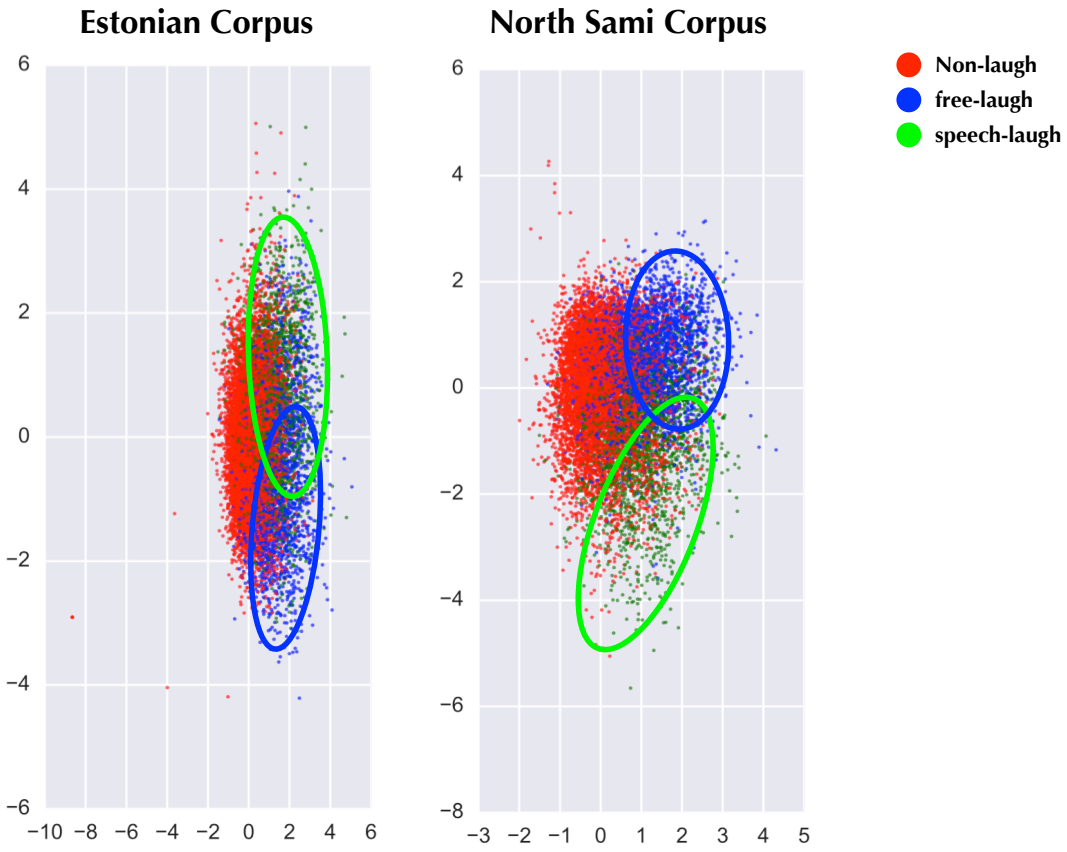

Figure 7: Applying LDA on MFCC features with context windows of 10 for both dataset, with marked laugh types

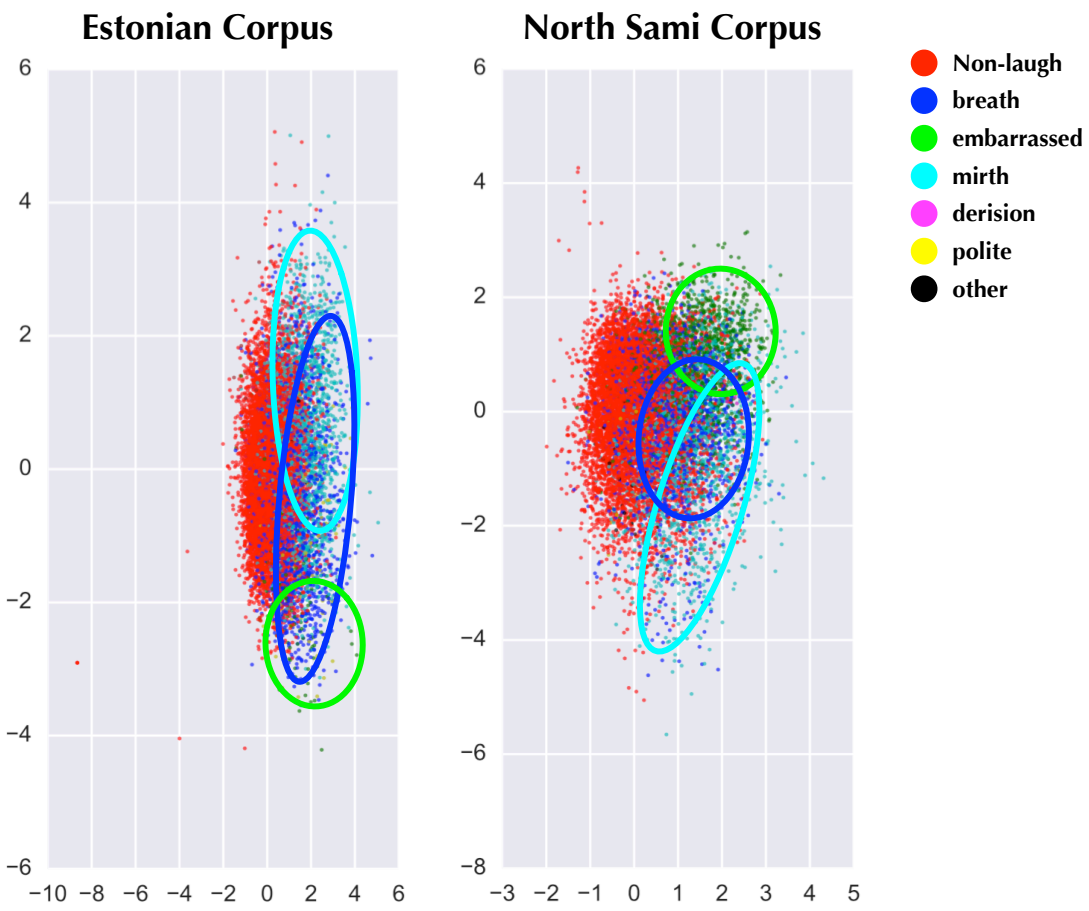

Figure 8: Applying LDA on MFCC features with context windows of 10 for both dataset, with marked emotional state

3. Acoustic system performance (binary classification)

First, we treat the problem as laugh detection, where we try to recognise laugh event from audio signal. All of our experiments uses MFCC features, unless otherwise specified, all model use 24 context windows. For deep learning systems, a rectifier function is applied on each hidden layer and the output layer use sigmoid function. Moreover, convolution neural network has pooling layer with pool_size of (2, 2) after each convolution with kernel size of (3, 3). Taking into account the fact that all results from our experiments are random variables, most of the experiments are run at least 3 times then we calculate the mean and standard deviation for each one. Table 1 shows the performance of different algorithms.

	Accuracy	Precision	Detection rate	F1
Logistic Regression	95.5	0	0	0
Linear SVM	60.3 ± 7.7	4.3 ± 0.5	41.7 ± 10.1	8
Deep learning (512-256)	96.7 ± 0.04	79.0 ± 2.2	34.7 ± 2.5	48.0 ± 2.2
Deep learning (512-256-128)	96.8 ± 0.1	74.7 ± 3.1	40.3 ± 2.6	52.3 ± 1.9
Convolution Neural Network (32-64)	96.8 ± 0.05	74.0 ± 3.7	41.3 ± 4.5	52.7 ± 2.6
Convolution Neural Network (32-64-64) context=80	97.1	76	47	58
Convolution Neural Network (32-64-64) context=100	96.94	64	63	63
Convolution Neural Network (32-64-64) context=200	97.75 ± 0.05	71.5 ± 2.5	62.2 ± 2.0	66.5 ± 0.5
Convolution Neural Network (32-64-64) context=250	97.6	60	73	66

Table 1: Performance of different system on binary classification task.

Since our dataset is biased, only 5.5% of data represent laugh event, accuracy is unreliable performance estimation. Precision represent positive predictive value, if our system is able to achieve 70% of precision, it means that when we predict an audio segment is laughing signal, we have 70% chance to make the right prediction. On the hand, detection rate (recall) is the amount of laugh events our system successfully detected (i.e. fraction of true laughing events that are retrieved). F1 score is the most reliable evaluation metrics, since it balances the precision and detection rate and is affected by a biased dataset.

From Table 1, we can see Logistic Regression and Linear SVM are strongly biased by the dataset, both system significantly misclassified non-laugh signal into laugh signal which result high detection rate but very low precision. The deep learning systems are gradually improved by increasing number of layer, however, our further experiments indicated that 3 layers is the optimal number of layers for given size of our dataset. By introducing convolutional neural network, we were able to slightly rise the performance by 0.4% (F1 score). Additionally, increasing number of context windows from 24 to 200 significantly improves the overall performance, our best system achieves 71.5% of precision with 62.2% detection rate and F1 equal to 66.5%. Since higher context length doesn't show improvement and requires more computational resource, 200 context windows are the optimal choice for the task.

Fig. 6 shows the detail performance of the best model, since the model are better at detection, it makes less mistakes in mis-classified non-laugh signal into laugh signal. However, there are many laughing samples were ignored by the model which illustrated at the bottom left corner of the figure.

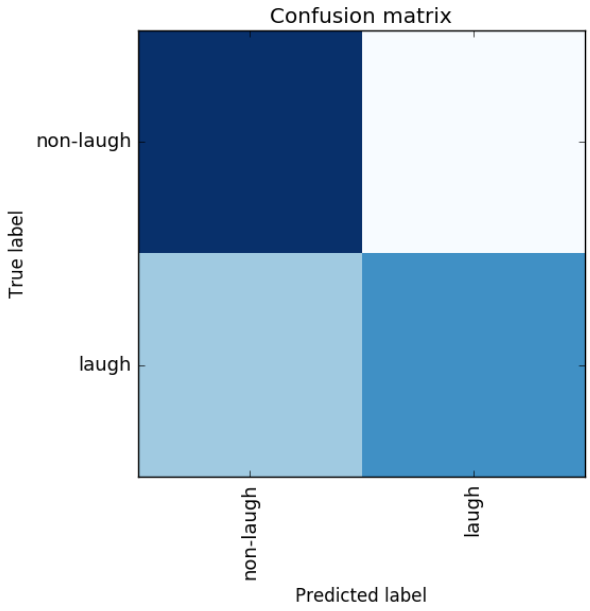

Figure 6: Confusion matrix on test data of the best model on binary classification task

4. Acoustic system performance (multi-classes classification)

In the multi-classes classification task, we train a convolutional neural network to discriminate each kind of laugh and non-laugh signal. The performance of the system are presented in Table 2, which is competitive to the best model in Section 3.

	Accuracy	Precision	Detection rate	F1
Convolution Neural Network (32-64-64) context=200	97.5 ± 0.4	62.8 ± 5.3	62.7 ± 8.3	96.0 ± 0.0

Table 2: Performance of best system on multi-classes classification task.

Since, the model simply ignores laughing kind with smaller number of data and focus on recognition the one with larger amount of data. The confusion matrix in Fig. 7 are biased. As "fl, m", "st, m", "fl, b" and "st, e"

acquires about 95% of laugh events, their performance are higher than other classes, and the system rarely misclassified their laugh signal into non-laugh signal.

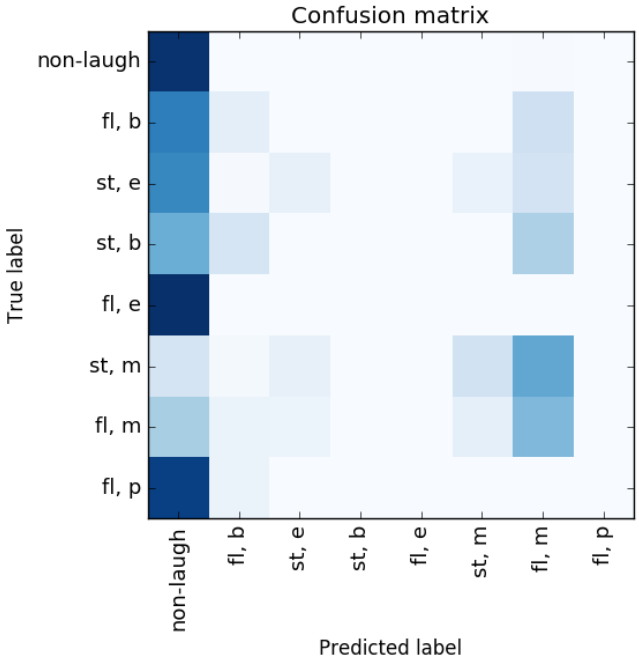

Figure 7: Confusion matrix on test data of the best model on multi-classes classification task

In conclusion, the binary classification task with convolutional neural network achieves more stable results on extreme skew dataset, with good performance on both detection and precision.