Differentiating Laughter Types via HMM/DNN and Probabilistic Sampling

Gábor Gosztolya[1,2(✉)], András Beke[3], and Tilda Neuberger[3]

[1] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
ggabor@inf.u-szeged.hu
[2] Department of Informatics, University of Szeged, Szeged, Hungary
[3] Research Institute for Linguistics of the Hungarian Academy of Sciences, Budapest, Hungary

Abstract. In human speech, laughter has a special role as an important non-verbal element, signaling a general positive affect and cooperative intent. However, laughter occurrences may be categorized into several sub-groups, each having a slightly or significantly different role in human conversation. It means that, besides automatically locating laughter events in human speech, it would be beneficial if we could automatically categorize them as well. In this study, we focus on laughter events occurring in Hungarian spontaneous conversations. First we use the manually annotated occurrence time segments, and the task is to simply determine the correct laughter type via Deep Neural Networks (DNNs). Secondly we seek to localize the laughter events as well, for which we utilize Hidden Markov Models. Detecting different laughter types also poses a challenge to DNNs due to the low number of training examples for specific types, but this can be handled using the technique of probabilistic sampling during frame-level DNN training.

Keywords: Laughter events · Deep Neural Networks · Hidden Markov Models · Probabilistic sampling

1 Introduction

Laughter is one of the most interesting and important aspects of complex human behaviour [25]. But why do humans have an ability to laugh, what is the evolutional purpose of laughter, and how did it develop during our evolution? To answer these questions, the function of laughter has to be analyzed from the perspective of human behaviour. It has been shown that there are many types

This study was partially funded by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413. Gábor Gosztolya was also supported by the Ministry of Human Capacities, Hungary (grant 20391-3/2018/FEKUSTRAT). András Beke was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

© Springer Nature Switzerland AG 2019
A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 122–132, 2019.
https://doi.org/10.1007/978-3-030-26061-3_13

of laughter depending on the approach used in the analysis. Based on the vocal-production mode, laughter can be realized as voiced or unvoiced, and there are intervals where a participant both speaks and laughs, known as speech-laughs (see e.g. [18]). Unvoiced laughter is acoustically similar to breathing. Voiced laughter was found to be a more relevant predictor of emotional involvement in speech than general laughter. Other types of laughter may be voiced song-like, unvoiced grunt-like, unvoiced snort-like and mixed sounds [3,14]. The types of laughter may be differentiated by considering the emotion of the speaker as well; for example hearty, amused, satirical and social laughs [23]. At least 23 types of laughter have been identified (hilarious, anxious, embarrassed, etc.), where each laughter type has its own social function [21].

More recently, there has been more interest in creating automatic classifiers that are able to differentiate laughter types based on acoustics, facial expressions and body movement features (e.g. [2,15,31]). The laughter detector developed by Campbell et al. [6] can automatically recognize four laughter types based on the speaker's emotion in Japanese (the identification rate is greater than 75%). The results of Galvan et al. also supported the possibility of automatical discrimination among five types of acted laughter: happiness, giddiness, excitement, embarrassment and hurtful [7]. In their study, automatic recognition based only on the vocal features achieved higher accuracy scores (70% correct recognition) than by using both facial and vocal features (60%) or just facial features alone (40%).

In a previous study ([22]), we discriminated laughter based on the perceived sound according to the identity and/or number of participants (test person, other person(s), both), and according to the connection between laughter and speech. We distinguished five types of laughter, namely

(i) single laughter (**S**): only the speaker's laughter can be heard,
(ii) overlapping laughter (**O**): two or more speakers' laughter occur at the same time,
(iii) laughter during the speech of others (**D**): the test person's laughter is heard while another participant or participants are speaking,
(iv) laughed speech (**P**): the speaker's laughter co-occurs with their own speech,
(v) mixed (**M**): a mixture of the previous three categories (ii) + (iii) + (iv).

These five categories of laughter may be associated with various functions in conversations. Single laughter may be a sincere emotional expression or reaction to one's own message or the others' message. Overlapping laughter may indicate a cooperative act. Laughter during the speech of others may be a sign of attention or a feedback to their message as a backchannel. Laughed speech may express the fact that the speaker intends to refine or moderate the content of their message. A mixed type of laughter has diverse functions in conversation.

Laughter – due to its various functions – contributes to the organisation of conversation. We can get closer to understanding the structure of the conversation by analysing laughter types. However, to do this, first they have to be located and identified. In this study we seek to automatically classify laughter segments as one of these five pre-defined categories; to do this, we borrow

Table 1. Some important properties of the different laughter types in the dataset used.

	Laughter type					All laughter types	All utterances
	Single	**O**ver-lapping	During other	Laughed **sp**eech	**M**ixed		
Total duration (m:ss)	2:12	2:13	4:17	1:52	1:27	12:01	147:36
% of duration	1.50%	1.50%	2.90%	1.26%	0.98%	8.14%	100.00%
Avg. duration (ms)	594	1087	937	1017	1887	930	—
Median duration (ms)	480	910	805	875	1620	740	—
No. of occurrences	223	122	274	110	46	775	—
% of occurrences	28.8%	15.7%	35.4%	14.2%	5.9%	100.0%	—
Frequency (1/s)	1.51	0.83	1.86	0.75	0.31	5.25	—

tools from Automatic Speech Recognition (ASR) such as acoustic feature sets and Deep Neural Networks (DNNs, [17]) for frame-level classification. To address both the *classification* and the *location* problems, in the second part of our study we combine the outputs of our frame-level DNNs with a Hidden Markov Model (HMM). However, as in laughter detection only a fraction of the training data corresponds to laughter, we shall use the sampling technique called *probabilistic sampling* [19] to assist frame-level DNN training.

2 The Recordings Used

Here, we used a part of the BEA Hungarian Spoken Language Database [9]. It is the largest speech database in Hungarian, which contains 260 h of material produced by 280 speakers (aged between 20 and 90 years), recorded in a sound-proof studio environment. In the present study we could use only the subset which had annotated laughter types at the time of writing, a total of 62 recordings of spontaneous conversations. The recordings lasted 148 min in total, from which we assigned 100 min (42 utterances) to the training set, while 20 and 27 min were assigned to the development set and the test set (10 recordings each). The segment boundaries of laughter segments were identified by human transcribers. Overall the total duration of laughter was 12 min, taking up 8.1% of all the utterances; of course, the different types of laughter were unevenly distributed.

Some main characteristics of the different laughter types in this dataset can be seen in Table 1. Unfortunately, the corpus we used is not very large, but it is typical in the area of laughter identification, especially if we can use only the utterances which have annotations about the types of laughter events. Surprisingly, the five types are roughly balanced when measured in total duration, the shortest sub-type (*Mixed*) taking up roughly 1% of the total playing time, and the most common one (*During others' speech*) comprised 2.9% of all the utterances. The main difference comes from the average duration and frequency of the types: the most frequently occurring laughter type was *Single*, but these laughter events were the shortest ones as well, while *Mixed* types occurred only once in three minutes of conversation, but then lasted for almost two seconds on average.

3 DNN Training by Probabilistic Sampling

For our experiments we borrowed techniques from Automatic Speech Recognition (ASR) such as Deep Neural Networks and Hidden Markov Models. Following standard ASR techniques, DNNs were used to provide a posterior probability estimate for each 10 ms for each utterance (i.e. for each *frame*). However, DNNs work best when they can be trained on hundreds or even thousands of hours of speech data (see e.g. [20]), and this amount is typically not available for laughter corpora. A further difference is that in ASR the classes are more-or-less uniformly present among the training frames, while in laughter detection only 4–8% of the duration corresponds to laughter, and the vast majority of training data belongs to the *non-laughter* class (i.e. other speech, silence and background noise). When we split the laughter class into several new classes, this class imbalance grows further.

The simplest solution for balancing the class distribution is to downsample the more frequent classes. This, however, results in data loss, hence it may also result in a drop in accuracy especially as our training set was quite small in the first place. A more refined solution is to *upsample* the rarer classes: we utilize the examples from these classes more frequently during training. A mathematically well-formulated upsampling strategy is the method called probabilistic sampling [19,29]. Probabilistic sampling selects the next training example following a two-step scheme. First we select a class according to some probability distribution, then we pick a training sample from the samples that belong to this class. For the first step, we assign the following probability to each class:

$$P(c_k) = \lambda \frac{1}{K} + (1 - \lambda) Prior(c_k), \tag{1}$$

where $Prior(c_k)$ is the prior probability of class c_k, K is the number of classes and $\lambda \in [0, 1]$ is a parameter. When $\lambda = 0$, the above formula returns the original class distribution, so probabilistic sampling will behave just as conventional sampling does. When $\lambda = 1$, we get a uniform distribution over the classes, so we get totally balanced samples with respect to class frequency. Selecting a value for λ between 0 and 1 allows us to linearly interpolate between the two distributions. According to our previous results, using probabilistic sampling can aid DNN training when the task is to detect laughter events [13] as well as other phenomena with rate occurrences such as filler events [12].

4 Classification Experiments

In the first series of experiments we just classify the laughter occurrences into one of the five types, relying on the manually annotated starting and ending points of the laughter segments. We simply trained our DNNs at the frame level and took the product of their output likelihoods, as in our previous studies we found that this approach worked quite well (see e.g. [11]). Following the results of preliminary tests, we divided the frame-level posterior estimates of the DNNs by the original class priors, which is common in HMM/DNN hybrids [4].

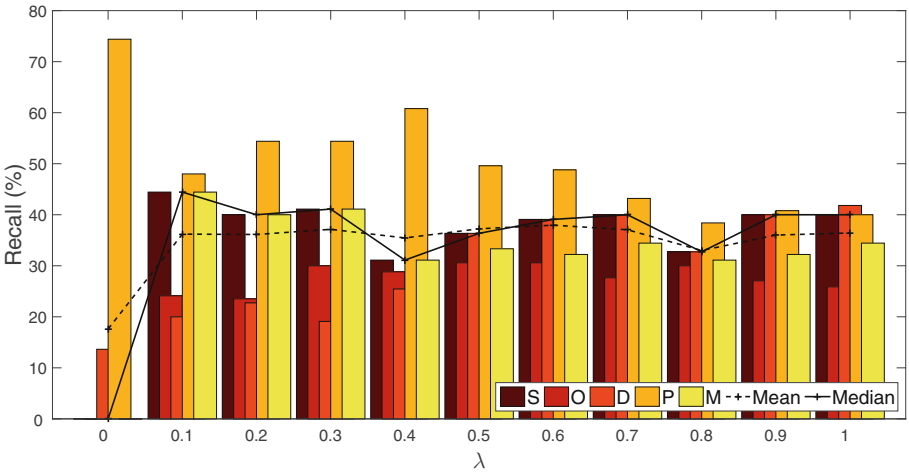

Fig. 1. Recall scores for the five laughter types on the development set as a function of λ.

Table 2. Some notable classification accuracy scores on the test set.

Classification method		Acc. (%)	Recalls (%)	
			Mean	Median
DNN with full sampling (baseline)		37.2	22.7	5.1
DNN + probabilistic sampling	$\lambda = 0.1$	35.3	**28.8**	**28.0**
	$\lambda = 0.6$	32.4	**27.2**	**26.1**
	$\lambda = 1.0$	30.1	**28.3**	**30.5**

4.1 DNN Parameters

We applied a DNN that had rectified linear units as hidden neurons [8,28] for frame-level classification. We used our custom DNN implementation [16], which achieved the best accuracy score published so far on the TIMIT database [27]. We employed DNNs with 5 hidden layers, each containing 256 rectified neurons. We applied the softmax function in the output layer. We used 40 mel filter bank energies as features along with first and second order derivatives, extracted using the HTK tool [30]. Training was performed on a sliding window containing 20 neighbouring frames from both sides, following the results of preliminary tests. Note that this sliding window size is quite large compared to ones used in speech recognition; but for laughter detection, using this many frames is clearly beneficial (see e.g. [5,11]).

4.2 Probabilistic Sampling

We evaluated the probabilistic sampling technique by varying the value of λ in the range $[0, 1]$ with a step size of 0.1. To reduce the effect of DNN random weight initialization, we trained five DNN models for each λ value; then we chose the value of λ based on the results obtained on the development set.

4.3 Evaluation

Since this was simply a classification task, we could have measured efficiency using the standard classification accuracy metric. However, it is well known that when class distribution is uneven, classification accuracy is biased towards the classes having more examples. Therefore we decided to calculate the recall of each laughter type. Afterwards, we aggregated the five recall values into one accuracy score via a simple arithmetic mean and median.

4.4 Results

Figure 1 shows the recall values got on the development set for all laughter types as a function of the λ parameter of probabilistic sampling. It is quite apparent that the values are not really consistent without applying probabilistic sampling (shown as $\lambda = 0$): actually no examples were classified as laughter types S, D and M. Using larger values for λ tends to balance the recall values of the five kinds of laughter, which is also reflected in the mean and median values. In our opinion, when the task is to identify the occurrences of distinct laughter sub-classes, the performance of an approach is more accurately described by the mean and even more so by the median of the recall values than traditional classification accuracy scores. Clearly, for values $\lambda \geq 0.5$ our approach works well for all laughter types, while it leads to a lower classification accuracy score.

Table 2 lists the accuracy, mean and median recall scores we got on the test set for some notable values of λ. (Values exceeding the baseline score are shown as **bold**.) Notice that the baseline case has the highest classification accuracy score (37.2%), but the low mean and especially the median recall value (5.1%) suggests a highly uneven behaviour. Overall, like that for the development set, all values of $\lambda \geq 0.1$ give a similar performance, which is significantly better than that for the baseline DNNs trained without probabilistic sampling.

5 Experiments with a Hidden Markov Model

Laughter (segment) classification is a simplified task in the sense that we rely on segment starting and ending points marked by human annotators. In the last part of our study we perform laughter *detection*, where, besides laughter types, we also have to find the *locations* of the different occurrences. We will do this by incorporating our likelihood values supported by DNNs into a Hidden Markov Model (HMM). In this set-up, the state transition probabilities of the HMM

practically correspond to a state-level bi-gram language model. Following the study of Salamin et al. [26], we calculated the model from statistics of the training set; the weight of this language model was determined on the development set, individually for the five DNN models trained.

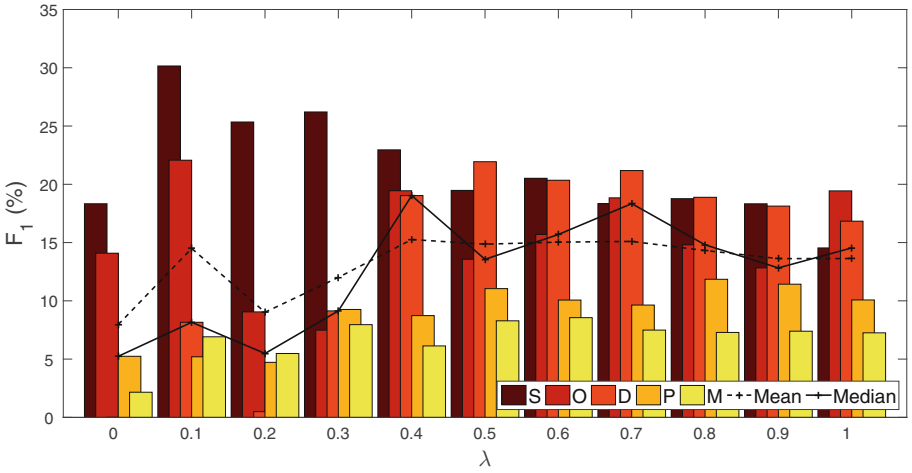

Fig. 2. Segment-level F_1 scores measured on the development set.

Table 3. Some notable segment-level F1 scores obtained on the test set.

	F_1(%)						
Sampling approach	**S**	**O**	**D**	**P**	**M**	Mean	Median
DNN (baseline)	18.3	14.1	0.0	5.2	2.2	8.0	5.2
DNN + prob. sampling, $\lambda = 0.4$	**23.0**	**19.5**	**19.0**	**8.7**	**6.1**	**15.3**	**19.0**
DNN + prob. sampling, $\lambda = 1.0$	14.5	**19.4**	**16.8**	**10.1**	**7.3**	**13.6**	**14.5**

5.1 Evaluation Metrics

We opted for the information retrieval (IR) metrics of *precision*, *recall* and their harmonic mean, *F-measure* (or F_1). To decide whether two occurrences of events (i.e. a laughter occurrence hypothesis returned by the HMM and one labeled by an annotator) match, there is no de facto standard in the literature. In this study we required that the two occurrences intersect (as in [10] and [24]), while their centre also had to be close to each other (within 500 ms, as in [1]). Furthermore, following the work of Salamin et al. [26], we calculated these metrics at the frame level as well. Since the optimal meta-parameters (λ and language model weight) may differ in the two (evaluation) approaches used, we set them independently.

5.2 Results

Figures 2 and 3 show the averaged F_1 scores got on the development set at the segment level and frame level, respectively. It can be seen that the smaller λ values ($\lambda \leq 0.3$ and $\lambda \leq 0.4$, at the segment and frame level, respectively) led to quite low F_1 values for some laughter types, while for larger λ parameters we had a more balanced behaviour. This is also reflected in the mean and median F_1 scores. At the segment level, optimality is achieved with $\lambda = 0.4$, while at the frame level it is with $\lambda = 0.1$ (mean) and with $\lambda = 0.5$ (median).

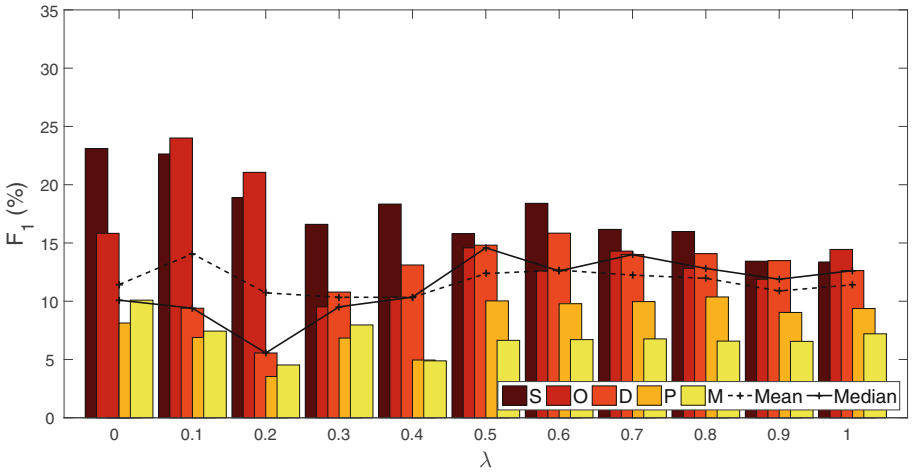

Fig. 3. Frame-level F_1 scores measured on the development set.

Table 4. Some notable frame-level F1 scores obtained on the test set.

Sampling approach	F_1(%)						
	S	**O**	**D**	**P**	**M**	Mean	Median
DNN (baseline)	23.1	15.8	0.0	8.1	10.1	11.4	10.1
DNN + prob. sampling, $\lambda = 0.1$	22.6	**24.0**	**9.4**	6.9	7.4	**14.1**	9.4
DNN + prob. sampling, $\lambda = 0.5$	15.8	14.6	**14.8**	**10.0**	6.7	**12.4**	**14.6**
DNN + prob. sampling, $\lambda = 1.0$	13.4	14.4	**12.6**	**9.4**	7.2	**11.4**	**12.6**

Overall, the F_1 scores seem to be somewhat low, even after applying probabilistic sampling. In our opinion, however, these are quite realistic scores, for two reasons. Firstly, even when we treat laughter as one class, we get F_1 values between 40 and 60% (see e.g. [10,13,26]), which is likely to be reduced further when we split the laughter class into several sub-classes. Secondly, recall that the laughter sub-types were defined based on the relation between the laughter

event and the other speaker's speech. This, combined with the large duration of laughter events, eventually leads to mixed laughter occurrences. For example, in an overlapping laughter event both the speakers are probably not laughing for the whole duration, but in some parts only one of them is (while the other speaks or remains silent). This, however, is quite hard to detect at the frame level.

Examining Tables 3 and 4 (containing the interesting F_1 scores obtained on the test set at the segment level and frame level, respectively), we see that the F_1 value of the Mixed laughter type is the lowest, which is probably due to the latter phenomenon. Overall, the F_1 values are more balanced for the different laughter types when using probabilistic sampling, and when we use the λ values found optimal on the development set, we get better results than either without probabilistic sampling or with uniform sampling (i.e. $\lambda = 1$). We got the highest frame-level mean F_1 value in the case where the mean was highest on the development set ($\lambda = 0.1$), and the same holds for the median ($\lambda = 0.5$). Overall, optimizing for the median led to a more balanced performance than optimizing for the mean, which led to a mixture or relatively high and low F_1 values.

6 Conclusions

In this study we sought to detect and identify multiple laughter types in Hungarian spontaneous conversations. We performed simple classification experiments and those where the location of laughter occurrences had to be determined as well. Overall, we found that the median of F_1 scores characterizes performance better than the arithmetic mean does, and the technique of probabilistic sampling aids the training of frame-level DNNs in the task of laughter sub-group classification, where the training data has a highly imbalanced class distribution.

References

1. NIST Spoken Term Detection 2006 Evaluation Plan (2006). http://www.nist.gov/speech/tests/std/docs/std06-evalplan-v10.pdf
2. Ayadi, M.E., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
3. Bachorowski, J.A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. J. Acoust. Soc. Am. **110**(3), 1581–1597 (2001)
4. Bourlard, H., Morgan, N.: Connectionist Speech Recognition - A Hybrid Approach. Kluwer Academic (1994)
5. Brueckner, R., Schuller, B.: Hierarchical neural networks and enhanced class posteriors for social signal classification. In: Proceedings of ASRU, pp. 362–367 (2013)
6. Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Proceedings of Interspeech, pp. 465–468, Lisbon, Portugal (2005)
7. Galvan, C., Manangan, D., Sanchez, M., Wong, J., Cu, J.: Audiovisual affect recognition in spontaneous Filipino laughter. In: Proceedings of KSE, pp. 266–271 (2011)

8. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier networks. In: Proceedings of AISTATS, pp. 315–323 (2011)
9. Gósy, M.: BEA: a multifunctional Hungarian spoken language database. Phonetician **105**(106), 50–61 (2012)
10. Gosztolya, G.: On evaluation metrics for social signal detection. In: Proceedings of Interspeech, pp. 2504–2508, Dresden, Germany, September 2015
11. Gosztolya, G., Beke, A., Neuberger, T., Tóth, L.: Laughter classification using Deep Rectifier Neural Networks with a minimal feature subset. Arch. Acoust. **41**(4), 669–682 (2016)
12. Gosztolya, G., Grósz, T., Tóth, L.: Social signal detection by probabilistic sampling DNN training. IEEE Trans. Affect. Comput. (2019, to appear)
13. Gosztolya, G., Grósz, T., Tóth, L., Beke, A., Neubergers, T.: Neurális hálók tanítása valószínűségi mintavételezéssel nevetések felismerésére. In: Proceedings of MSZNY, pp. 136–145, Szeged, Hungary (2017). (in Hungarian)
14. Grammer, K., Eibl-Eibesfeldt, I.: The ritualisation of laughter, Chapter 10. In: Natürlichkeit der Sprache und der Kultur: Acta colloquii, pp. 192–214, Brockmeyer (1990)
15. Griffin, H.J., et al.: Laughter type recognition from whole body motion. In: Proceedings of ACII, pp. 349–355 (2013)
16. Tóth, L., Grósz, T.: A comparison of deep neural network training methods for large vocabulary speech recognition. In: Habernal, I., Matoušek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 36–43. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40585-3_6
17. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
18. Laskowski, K.: Contrasting emotion-bearing laughter types in multi participant vocal activity detection for meetings. In: Proceedings of ICASSP, pp. 4765–4768 (2009)
19. Lawrence, S., Burns, I., Back, A., Tsoi, A.C., Giles, C.L.: Neural network classification and prior class probabilities. In: Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 1524, pp. 299–313. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-49430-8_15
20. McDermott, E., Heigold, G., Moreno, P., Senior, A., Bacchiani, M.: Asynchronous stochastic optimization for sequence training of Deep Neural Networks: towards big data. In: Proceedings of Interspeech, pp. 1224–1228, September 2014
21. McKeown, G., Cowie, R., Curran, W., Ruch, W., Douglas-Cowie, E.: Ilhaire laughter database. In: Proceedings of LREC, pp. 32–35 (2012)
22. Neuberger, T., Beke, A.: Automatic laughter detection in Hungarian spontaneous speech using GMM/ANN hybrid method. In: Proceedings of SJUSK Conference on Contemporary Speech Habits, pp. 1–13 (2013)
23. Ohara, R.: Analysis of a laughing voice and the method of laughter in dialogue speech. Master's thesis, Nara Institute of Science and Technology, Ikoma, Japan (2004)
24. Pokorny, F.B., et al.: Manual versus automated: the challenging routine of infant vocalisation segmentation in home videos to study neuro(mal)development. In: Proceedings of Interspeech, San Francisco, CA, USA, pp. 2997–3001, September 2016
25. Ross, M.D., Owren, M.J., Zimmermann, E.: The evolution of laughter in great apes and humans. Commun. Integr. Biol. **3**(2), 191–194 (2010)

26. Salamin, H., Polychroniou, A., Vinciarelli, A.: Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In: Proceedings of SMC, pp. 4282–4287 (2013)
27. Tóth, L.: Phone recognition with hierarchical Convolutional Deep Maxout Networks. EURASIP J. Audio Speech Music Process. **2015**(25), 1–13 (2015)
28. Tóth, L.: Phone recognition with deep sparse rectifier neural networks. In: Proceedings of ICASSP, pp. 6985–6989 (2013)
29. Tóth, L., Kocsor, A.: Training HMM/ANN hybrid speech recognizers by probabilistic sampling. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3696, pp. 597–603. Springer, Heidelberg (2005). https://doi.org/10.1007/11550822_93
30. Young, S., et al.: The HTK Book. Cambridge University Engineering Department, Cambridge (2006)
31. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)