# The Influence of Pleasant and Unpleasant Odours on the Acoustics of Speech

Maurice Gerczuk[1], Anton Batliner[1], Shahin Amiriparian[1], Andreas Triantafyllopoulos[1], Franziska Heyne[2], Marie Klockow[2], Thomas Hummel[2], and Björn W. Schuller[1,3]

*Affiliation 1:* Chair EIHW, University of Augsburg, Augsburg, Germany,
{maurice.gerczuk,anton.batliner,shahin.amiriparian,andreas.triantafyllopoulus,bjoern.shuller}@uni-a.de
*Affiliation 2:* Smell & Taste Clinic, Dept. of ORL, TU Dresden, Germany,
{franziska.heyne,marie.klockow,thomas.hummel}@tu-dresden.de
*Affiliation 3:* Group on Language, Audio, & Music, Imperial College, UK

***Abstract*—Olfaction, i. e., the sense of smell is referred to as the 'emotional sense', as it has been shown to elicit affective responses. Yet, its influence on speech production has not been investigated. In this paper, we introduce a novel speech-based smell recognition approach, drawing from the fields of speech emotion recognition and personalised machine learning. In particular, we collected a corpus of 40 female speakers reading 2 short stories while either no scent, unpleasant odour (fish), or pleasant odour (peach) is applied through a nose clip. Further, we present a machine learning pipeline for the extraction of data representations, model training, and personalisation of the trained models. In a leave-one-speaker-out cross-validation, our best models trained on state-of-the-art wav2vec features achieve a classification rate of 68 % when distinguishing between speech produced under the influence of negative scent and no applied scent. In addition, we highlight the importance of personalisation approaches, showing that a speaker-based feature normalisation substantially improves performance across the evaluated experiments. In summary, the presented results indicate that odours have a weak, but measurable effect on the acoustics of speech.**

***Keywords*—digital health, computational paralinguistics, scent, odour, emotion, speech**

## I. Introduction

Apart from its major role in food intake, olfaction, i. e., the sense of smell, is associated with emotions and hedonics, i. e., the perception of pleasant and unpleasant sensations [1]. Odours can evoke emotional memories [2], improve mood [3] or reduce anger [4]. Scent has further been shown to play a role in social interactions, such as reproductive behaviour and emotional contagion [1]. Moreover, olfactory dysfunction and impairment of odour memory have been associated with major depressive disorder [5]. Most (experimental) studies concerning the relationship of emotional states and olfaction so far deal with cross-modality and perception of scents; all in all, they agree by concluding that positive scents are associated with brightness and higher pitch and by that, positive valence in emotions [6–12]. Given these characteristics, olfaction is an interesting topic to approach from an affective computing point of view. Sabiniewicz et al. [13] utilised automatic recognition of facial expressions to measure changes in emotional state according to four basic emotions – anger, happiness, sadness, and surprise. In the context of automatic emotion recognition, speech has long played an important role [14, 15] as affective states are mirrored in the perception of the acoustics of speech [16]. From this perspective, the influence of scent on affective states might affect the production of speech as well. To the best of our knowledge, the study by Millot et al. [17] is the only one addressing the influence of scents on speech production. They conclude that the pitch of the voice had a positive correlation with the pleasantness of scent, i. e., pitch was higher in the pleasant versus the unpleasant odour condition.

Recently, personalisation approaches, i. e., strategies that aim to efficiently adapt machine learning models to individuals, have been explored in the context of digital health applications, such as the prediction of mood and stress from smartphone-collected data [18]. When it comes to low-data speech analysis settings, Triantafyllopoulos et al. [19] showed the significance of speaker-based feature normalisation for the classification of pre- and post-treatment speech of patients with chronic obstructive pulmonary disease. As a person's reaction to a specific scent is informed by previous experiences with the corresponding odorous item and thus partly driven by a learning process [20, 21], it can be complex and highly subjective [13]. Consequently, we assume that scent-induced variations in speech production will be hard to generalise across subjects and therefore employ the same paradigm.

In the present study, we want to investigate whether smelling pleasant or unpleasant scents – in relation to the neutral control scenario – leaves traces in the acoustics of speakers that can further be picked up by an automatic paralinguistic analysis in the form of personalised machine learning.

## II. DATASET AND EXPERIMENTAL DESIGN

A total of 40 healthy female speakers (German natives) with a subjective normal sense of smell, aged 19 to 39 years (median: 25 years), were recruited. They were mostly students and were renumerated/paid for their participation. The speakers were randomly assigned to two groups, G1 and G2, and had to read two short stories in German and retell a picture book in four rounds R1-R4. Each round lasted around five minutes. The scents applied were water (neutral), peach (positive), and fish (negative). Their order is different for the two groups:

- Group 1: R1 water, R2 peach, R3 fish, R4 water
- Group 2: R1 water, R2 fish, R3 peach, R4 water

After each round, there was a pause of approximately 5-10 minutes; in this time, the speakers answered a questionnaire regarding the applied scent. The recordings took place in a quiet office. The scents were applied with the help of a u-shaped plastic clip (aspura clip® mini inhalator). Recordings were done with the Smartphone HUAWEI Pro lite 2017, Model PRA-LX1, with RecForge II Pro – Audio Recorder (44 kHz, Mono), using the Rode smartLav + and the splitter Rode SC6 special adapter. As we want to employ forced alignment in the present study, we only use the two read stories and disregard the retelling for the moment.

We consider the dependent variable, scent, with three classes, neutral (water), positive (peach), and negative (fish). Furthermore, there are two (possibly interacting) intervening variables. We have to account, firstly, for habituating to reading the same stories four times in a short time, and, secondly, for the probability that speakers might be more or less sensitive towards the scents applied. Therefore, we carefully select the set of experimental configurations not to confound habituation for scent-induced variations in speech production and arrive at four binary classification setups that can further be summarised in two groups. First, classifying positive vs negative scent (peach vs fish), and, secondly, classifying non-neutral scent against neutral scent. In the second group, we evaluate classifying the negative and positive scents against neutral scent individually (fish vs neutral, peach vs neutral) and combined (scent vs neutral).

## III. EXPERIMENTAL SETUP

Our experimental pipeline consists of four consecutive steps. The speech recordings are automatically segmented into linguistic phrases via forced alignment after which we extract paralinguistic representations from these segments and apply one of two different normalisation strategies. Finally, we train linear Support Vector Machines (SVMs) on the data to perform the classification of applied scent.

### *A. Segmentation*

As a first preprocessing step, we segment the speech recordings into prosodic phrases utilising forced alignment to the transcriptions with the Munich Automatic Segmentation System (MAUS) [22]. The first story (A) contains a total of 13 phrases while we separate story B into 24 phrases. In total, this results in 6080 phrases – both stories are read in each of the 4 rounds by each of the 40 speakers.

### *B. Features*

We evaluate and compare the efficacy of two audio data representations. The eGeMAPs [23] set of audio functionals is handcrafted for the task of Speech Emotion Recognition and provides interpretability due to its comparatively small size (88 numeric features) whereas our second choice, deep features extracted from a pre-trained wav2vec2 [24] model, sacrifice post-hoc feature analysis in favour of modelling phonological information – representing the current state-of-the-art for automatic speech recognition. We extract eGeMAPs and wav2vec2 features with openSMILE [25] and huggingsound [26], respectively.

### *C. Normalisation*

Two different z-score normalisation strategies are explored, differing in the subsets of data over which the statistics are computed.

**Global**: as a baseline, we perform a global feature normalisation of the data, i. e., we compute feature statistics on the training data of each fold and apply them to normalise each partition.

**Subject-level**: The second procedure is based on computing (and applying) mean and standard deviation normalisation for each subject independently, using samples from the second to fourth rounds. We do not include the first round in the data, as the habituation effect is especially pronounced when comparing this round to all following rounds.

### *D. Classification and Evaluation*

To perform the actual classifications of applied scent and habituation from speech recordings of participants, we train linear support vector machines (SVMs) on the extracted and normalised features, optimising the cost parameter on the validation data (cf. Section III-D). As our dataset is quite small, we opt for a Leave-One-Speaker-Out (LOSO) cross-validation setup, i. e., in each fold, we leave out one speaker's samples as testing data and train an SVM on the remaining speakers' data. In this way, we generate exactly one prediction for every sample in our dataset. We evaluate our models' performance based on the unweighted average recall (UAR) computed from these predictions and the ground truth. Note that, in our case, UAR is the same as traditional accuracy in all, but one experimental setup, i.e., the number of samples in the classes are equal, unless we consider negative and positive odour together.

As our baseline unit of analysis are individual phrases of each of the two read stories, we have several data points for every unique combination of subject and round. We suspect that the influence of specific smells or habituation on a subject's voice might vary in the course of reading the short stories, e. g. because the reader habituates to the smell. To investigate this, we aggregate model predictions for each phrase into larger units via a majority vote to receive a prediction for each subject and round.

TABLE I. Phrase- and subject-level UAR% (phrase/subject) for evaluated scent experiments. Here, subject-level normalisation is done on the basis of rounds 2-4 for each speaker.

| Features | Normalisation | peach vs fish | scent vs neutral | peach vs neutral | fish vs neutral |
|---|---|---|---|---|---|
| wav2vec | Global | 50/50 | 53/55 | 52/56 | 52/54 |
| | | 49/**55** | 53/57 | 52/**57** | 53/59 |
| | Subject | 50/54 | 52/57 | 52/52 | 52/55 |
| | | 51/50 | 54/**67** | 53/**57** | 57/**68** |

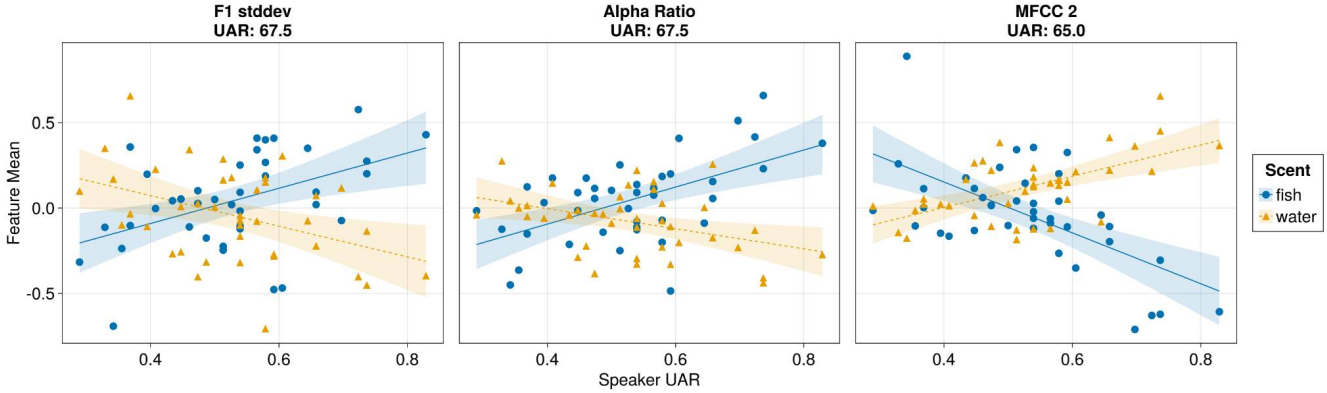


Fig. 1: Relationship between per-speaker UAR and mean feature values when classifying fish vs neutral scent. The most important features (based on classification performance) are the standard deviation of the first formant (F1 stddev), the alpha ratio, and the second MFCC (MFCC 2). UARs are plotted against the corresponding mean feature values per speaker and round, i. e., there is one data point for every unique combination of speaker and applied scent.

## IV. Results

We present the results achieved in our experiments in Table I, grouped by feature representation (eGeMAPs and wav2vec), normalisation strategy (global- and subject-level), and experiment setup. Additionally, we perform an analysis of some individual audio functionals.

### A. *Classification*

In general, we observe low performance when classifying negative (fish) vs positive scent (peach) from the speech of participants, only slightly above chance level in some cases. When we task our models with distinguishing between unaffected speech (neutral scent) and speech recorded with scent applied, more pronounced variations seem to emerge, leading to higher performance. Best results can be achieved when classifying negative against neutral scent, reaching a UAR of 68 % with speaker-normalised wav2vec features and after majority voting. Comparatively, the detection of positive scent falls behind by roughly 10 % with the same normalisation and features. However, this discrepancy cannot be found when using eGeMAPs where performance is about the same. As a handcrafted set of audio functionals, eGeMAPs was specifically designed to include features relevant to speech emotion recognition [23] whereas wav2vec's training and downstream applications are targeted more towards speech recognition and thus capture prosodic information. In this way, our results suggest that negative scent affects articulation more strongly than positive scent.

The effect of our normalisation-based personalisation approach is mostly observed when we look at the results after applying a majority vote to all predictions of one speaker in a particular round. Here, the models trained on wav2vec features to classify neutral vs non-neutral scent benefit the most with performance increases of up 13 %.

### B. *Feature interpretation*

Further, we perform an analysis of some individual features that have a large impact on the decisions of our models regarding their relationship to applied scent. As we found that classification works in the case of fish vs water, we focus on this setting. We utilise SHAP (SHapley Additive exPlanations) [27] to identify each feature's importance for the classification when used in conjunction with the rest of the features. Moreover, we consider each feature in isolation to train a model according to the experimental setup described before. We then choose the three best-performing features and plot the speaker-normalised feature means against the respective per-speaker classification performance (measured in UAR).

This visualisation (cf. Figure 1) allows us to compare how the features change under the influence of negative scent and further analyse how these changes interact with the performance obtained during classification. The most discriminating features are found with the standard deviation of the first formant (F1 standard deviation), the alpha ratio, and the second Mel-frequency cepstral coefficient (MFCC 2). We observe higher deviations in the frequency of the first formant and a higher alpha ratio when a negative scent (fish) is applied during the readings. Generally, high variance in formant frequencies can indicate voice instabilities [28] and increases in alpha ratio have been associated with fatigue [29]. Lastly, a decrease in the average of the second MFCC across voiced segments has been observed in depressed individuals where reduced muscular tension leads to a more closed mouth position [30]. Note that all these effects are not very pronounced, see UAR reported in Fig. 1. However, they all correspond to characteristics of negative traits and states such as voice instabilities, fatigue, depression -- and obviously, smelling unpleasant odours, i.e., fish.

## V. Conclusion

In this study, we investigated the effects of negative and positive scents on the production of speech in a database of 40 female speakers. We applied two machine learning paradigms originating from the fields of speech emotion recognition and general speech recognition in four carefully chosen experimental setups. Our best approaches, based on state-of-the-art wav2vec features could achieve the best classification rate of 68% UAR when tasked with distinguishing speech produced under the influence of fish odour from recordings where no scent was applied. As reactions to scent are highly subjective, our applied personalisation strategy was further shown to lead to substantial performance gains. In summary, the presented results indicate that odours have a weak, but measurable effect on speech. In the future, more involved personalisation strategies should be explored. A good fit could be found with enrolment-based approaches, utilising samples from neutral rounds to adapt a neural network to each speaker [31]. Further, low-resource speech processing frameworks such as DEEPSPECTRUMLITE [32] can be used for the real-time application of a speech-based scent classifier on embedded devices.

## Acknowledgment

This research was partially supported by the Deutsche Forschungsgemeinschaft (DFG) under grant agreements No. 421613952 (ParaStiChaD) and No. 442218748 (AUDI0NOMOUS). We would like to thank aspUraclip GmbH, Mittelstraße 7, D-12529 Berlin-Schönefeld, Germany, for providing the aspiraclip® mini inhalators.

## References

[1] R. J. Stevenson, “An Initial Evaluation of the Functions of Human Olfaction,” Chemical Senses, vol. 35, no. 1, pp. 3–20, Jan. 2010, ISSN: 0379-864X, 1464-3553. DOI: 10.1093/chemse/bjp083.

[2] M. Larsson and J. Willander, “Autobiographical Odor Memory,” Annals of the New York Academy of Sciences, vol. 1170, no. 1, pp. 318–323, 2009, ISSN: 1749-6632. DOI: 10 . 1111 / j . 1749 - 6632 . 2009.03934.x.

[3] S. C. Knasko, “Ambient odor’s effect on creativity, mood, and perceived health,” Chemical Senses, vol. 17, no. 1, pp. 27–35, 1992, ISSN: 0379-864X, 1464-3553. DOI: 10.1093/chemse/17.1.27.

[4] A. Rétiveau, E. C. Iv, and G. Milliken, “Common and Specific Effects of Fine Fragrances on the Mood of Women,” Journal of Sensory Studies, vol. 19, no. 5, pp. 373–394, 2004, ISSN: 1745-459X. DOI: 10.1111/j.1745-459x.2004.102803.x.

[5] G. M. Zucco and F. Bollini, “Odour recognition memory and odour identification in patients with mild and severe major depressive disorders,” Psychiatry Research, vol. 190, no. 2-3, pp. 217–220, Dec. 2011, ISSN: 0165-1781. DOI: 10.1016/j.psychres.2011.08.025.

[6] E. Hornbostel, “Über Geruchshelligkeit,” Pflüger, Arch., vol. 227, pp. 517–538, 1931.

[7] K. Belkin, R. Martin, S. Kemp, and A. Gilbert, “Auditory Pitch as a Perceptual Analogue to Odor Quality,” Psychological Science, vol. 8, pp. 340–342, 1997.

[8] S. Kemp and A. Gilbert, “Odor intensity and color lightness are correlated sensory dimensions,” Am J Psychol., vol. 110, pp. 35–46, 1997.

[9] C. Velasco, D. Balboa, F. Marmolejo-Ramos, and C. Spence, “Cross-modal effect of music and odor pleasantness on olfactory quality perception,” Frontiers in Psychology, vol. 5, pp. 1332, 1–9, 2014.

[10] S. T. Glass and E. Heuberger, “Effects of a Pleasant Natural Odor on Mood: No Influence of Age,” Natural Product Communications, vol. 11, pp. 1555–1559, 2016.

[11] I. Kontaris, B. S. East, and D. A. Wilson, “Behavioral and Neurobiological Convergence of Odor, Mood and Emotion: A Review,” Frontiers in Behavioral Neuroscience, vol. 14, pp. 1–15, 2020.

[12] R. J. Ward, S. M. Wuerger, and A. Marshall, “Smelling Sensations: Olfactory Crossmodal Correspondences,” Journal of Perceptual Imaging, vol. 5, pp. 000402-1-000402–12, 2022.

[13] A. Sabiniewicz, F. Heyne, and T. Hummel, “Odors modify emotional responses,” Flavour and Fragrance Journal, vol. 36, no. 2, pp. 256–263, 2021, ISSN: 1099-1026. DOI: 10.1002/ffj.3640.

[14] S. Amiriparian, “Deep representation learning techniques for audio signal processing,” Ph.D. dissertation, Technische Universität München, 2019.

[15] M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, “Emonet: A transfer learning framework for multi-corpus speech emotion recognition,” IEEE Transactions on Affective Computing, 2021.

[16] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” Communications of the ACM, vol. 61, no. 5, pp. 90–99, Apr. 2018, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3129340.

[17] J. Millot and G. Brand, “Effects of pleasant and unpleasant ambient odors on human voice pitch,” Neuroscience Letters, vol. 297, pp. 61–63, 2001.

[18] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, “Personalized Multitask Learning for Predicting Tomorrow’s Mood, Stress, and Health,” IEEE Transactions on Affective Computing, vol. 11, no. 2, pp. 200–213, Apr. 2020, ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2017.2784832.

[19] A. Triantafyllopoulos, M. Fendler, A. Batliner, M. Gerczuk, S. Amiriparian, T. Berghaus, and B. W. Schuller, “Distinguishing between pre- and post-treatment in the speech of patients with chronic obstructive pulmonary disease,” in Proc. Interspeech 2022, 2022, pp. 3623–3627. DOI: 10.21437/Interspeech.2022-10333.

[20] H. Lapid, S. Shushan, A. Plotkin, H. Voet, Y. Roth, T. Hummel, E. Schneidman, and N. Sobel, “Neural activity at the human olfactory epithelium reflects olfactory perception,” Nature neuroscience, vol. 14, pp. 1455–61, Sep. 2011. DOI: 10.1038/nn.2926.

[21] T. Engen, Odor sensation and memory. Greenwood Publishing Group, 1991.

[22] F. Schiel, “Automatic Phonetic Transcription of Non-Prompted Speech,” in Proc. of the ICPhS, San Francisco, Aug. 1999, pp. 607–610.

[23] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” IEEE transactions on affective computing, vol. 7, no. 2, pp. 190–202, 2015.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460.

[25] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in Proceedings of the International Conference on Multimedia - MM ’10, Firenze, Italy: ACM Press, 2010, p. 1459, ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246.

[26] J. Grosman, HuggingSound: A toolkit for speech-related tasks based on Hugging Face’s tools, https://github.com/jonatasgrosman/huggingsound, 2022.

[27] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in neural information processing systems, vol. 30, pp. 4768–4777, 2017.

[28] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, “An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices,” BioMed Research International, vol. 2013, e758731, Oct. 2013, ISSN: 2314-6133. DOI: 10.1155/2013/758731.

[29] L. Rantala, L. Paavola, P. Körkkö, and E. Vilkman, “Working-day effects on the spectral characteristics of teaching voice,” Folia Phoniatrica et Logopaedica, vol. 50, no. 4, pp. 205–211, 1998.

[30] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, “Automatic modelling of depressed speech: Relevant features and relevance of gender,” in Celebrating the diversity of spoken languages: 15th annual conference of the International Speech Communication Association (INTERSPEECH 2014), Singapore, 14 - 18 September 2014, H. Li, Ed., 2014, ISBN: 9781634394352.

[31] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, “Deep speaker conditioning for speech emotion recognition,” in 2021 IEEE International Conference on Multimedia and Expo (ICME), Jul. 2021, pp. 1–6. DOI: 10.1109/ICME51207.2021.9428217.

[32] S. Amiriparian, T. Hübner, V. Karas, M. Gerczuk, S. Ottl, and B. W. Schuller, “DeepSpectrumLite: A Power-Efficient Transfer Learning