Combining Hierarchical Classification with Frequency Weighting for the Recognition of Eating Conditions

Johannes Wagner, Andreas Seiderer, Florian Lingenfelser, Elisabeth André

University of Augsburg, Lab for Human Centered Multimedia, Germany

[wagner,seiderer,lingenfelser,andre]@hcm-lab.de

Abstract

Though parents regularly remind their children not to do so, talking while eating is a typical everyday situation automatic speech analysis systems should be able to deal with. The *Paralinguistic Eating Condition (EC) Challenge* at INTERSPEECH 2015 sets the task to classify whether a speaker is eating or not, and if so, which type of food the speaker is currently tasting. The approach we follow in this paper is rather unusual: instead of suppressing the influence of noise to enhance the intelligibility of a spoken message, we try to emphasize the noisy parts of the spectrum to improve the recognition of food classes. To allow for a fine-grained adaption to the characteristic spectrum of single food types we adopt a hierarchical tree structure and decompose the classification task into a sequence of binary decisions. At each node we apply frequency-dependent weighting to tune the spectrum to the involved target classes. With our approach we are able to improve results in a 7-class recognition problem (6 types of food and no food) by more than 7% on the training set (using leave-one-eater-out cross validation) and 4% on the test set, respectively.

Index Terms: Computational Paralinguistics, Hierarchical Classification, Eating Condition

1. Introduction

The *Eating Condition (EC) Sub-Challenge* as part of the INTERSPEECH 2015 *Computational Paralinguistic Challenge* (COMPARE) [1] sets the task to classify whether a speaker is eating or not, and if so, which type of food the speaker is currently tasting [2]. At a first glance the task may appear rather specific and exotic, yet it is just a typical everyday situation automatic speech analysis systems should be able to deal with. The interesting aspect of this particular challenge is that it brings those parts of a speech signal into focus, which usually are treated as noise. Quite some work has been carried out to make speech recognition systems robust against background noise, like the driving noise inside a car [3] or environmental noise in a cafeteria [4]. Techniques to remove the effect of noise include spectral subtraction [5, 6], feature enhancement [7], and masking [8]. Learning to recognize and distinguish specific types of noise, however, will allow for more sophisticated ways to handle it, for instance, by model adaptation [9]. To this end, the approach we follow in this paper is rather unusual: instead of suppressing the influence of noise to improve the intelligibility of a spoken message like in [10], we try to emphasize the noisy parts of the spectrum. To allow for a fine-grained adaption to the characteristic spectrum of single food types we adopt a hierarchical tree structure and decompose the classification task into a sequence of binary decisions. At each node we apply frequency-dependent weighting to optimise the spectrum to the involved target classes. Though the approach proposed in this paper is not limited to the recognition of eating conditions, the provided data set is well-suited to investigate the potential of the approach. For one reason, because the food types featured in the corpus resemble each other at different degrees, which suits hierarchical classification. For another reason, since the individual consistence of food leads to characteristic frequency spectrums. For instance, Dacremont [11] reports that crispy food generates high pitched sounds above 5 kHz, whereas crunchy food generates low pitched sounds with a characteristic peak on frequency range 1.25 to 2 kHz. To some extent, the approach we propose in this paper carries on an earlier study in which we tried to optimize a sequence of cascading classifiers via feature selection [12]. However, the study at hand adopts a more generic tree structure and investigates the effect of frequency weighting instead of feature selection. It should be noted that apart from automatic speech recognition, the detection of food sounds is relevant for other areas of application, too, e. g. activity recognition [13, 14]. A thorough overview is given in [2].

	NO No_Food	**AP** Apple	**HA** Haribo	**NE** Nectarine	**BA** Banana	**CR** Crisp	**BI** Biscuit	**%**
NO	127	1	2	3	5	1	1	**90.7**
AP	5	69	11	29	8	8	10	**49.3**
HA	2	10	70	13	24	0	0	**58.8**
NE	1	33	13	45	31	4	6	**33.8**
BA	5	11	22	27	73	0	2	**52.1**
CR	1	7	1	4	1	106	20	**75.7**
BI	1	18	0	5	3	22	84	**63.2**
Ø								**60.5**

Table 1: Confusion matrix for a 7-way classification of eating conditions (6 types of food and no food). Most errors occur within two subsets ({Apple,Haribo,Nectarine,Banana} and {Crisp,Biscuit}). According cells are highlighted in gray.

2. Baseline

First, a baseline for upcoming experiments needs to be established. The data set provided by the challenge organizers contains read and spontaneous speech of 30 subjects while eating one of six types of food (Apple, Haribo, Nectarine, Banana, Crisp, Biscuit), as well as, clean samples (No_Food). The task is to assign a tested file to one of the 7 classes. A detailed description of the data set is found in [2]. Throughout the paper recognition rates will be reported as unweighted average recall (UAR), which is the standard measure of the INTERSPEECH Computational Paralinguistics Challenge series [15]. To tell wins and losses at a glance results are given relative to a base-

line we will introduce in a moment. For classification we apply linear kernel Support Vector Machines (SVM) with default settings as provided by the popular libSVM library (v3.20) [16] and extract the ComPareE feature set (6'373 features) [1] with OpenSmile (v2.1) [17, 18]. As suggested by the challenge organizers leave-one-eater-out (LOSO) cross-validation is used. We stick to this configuration throughout the experiments and no further optimizations such as parameter fine tuning or feature selection will be applied. Table 1 shows the confusion matrix for a 7-way classification of eating conditions (6 types of food and no food) applied on the training set. The UAR of 60.5% will be used as baseline for the following experiments (a comparison with the official challenge baseline will be included in the final discussion).

3. Methodology

Let us now introduce the methods we have investigated to improve classification of food types.

3.1. Hierarchical Classification

The last column of the confusion matrix in Table 1 lists individual recognition rates for the target classes. The values reveal that food types are recognized with varying degrees of success. The recall value of Crisp, for instance, is more than twice as high as that of Nectarine. The matrix also shows that most errors occur within two subsets {Apple, Haribo, Nectarine, Banana} and {Crisp, Biscuit}. It seems natural to assume that tasting food of the same subset produces a similar sound, which makes separation more challenging. To account for this and to simplify the classification task at first, related classes can be grouped. Afterwards specialized classifiers trained with the samples of according subsets are consulted to refine the decision. In other words, we decompose a complex classification problem into a sequence of simpler decisions. A suited structure to represent such a system is a hierarchical tree structure. At each node of the tree an individual classifier is installed to decide which branch should be followed. Classes in the remaining branches are sorted out. This is repeated until a leaf is reached. If the leaf holds a single class it is assigned as the winning class, otherwise a final classifier is consulted to derive a decision between remaining candidates. In our experiments we adopt a binary tree structure, i. e. at inner nodes classes are split in two disjunct subgroups.

3.2. Frequency Weighting

Figure 1 plots the mean spectrum of the six food classes. Though the graphs share a general course, we can also perceive individual differences. The cracking noise of Crisp, for instance, is reflected in the high amplitude for frequencies above 4 kHz, whereas Banana appears to have a low amplitude in this region due to the soft pulp causing smacking sounds. To improve audibility of a food type we can emphasize its dominant frequencies in the spectrum and suppress frequencies in remaining areas. As a matter of fact, this should lead to a more distinctive feature set and make it easier for a classifier to distinguish according sounds. In speech recognition, for instance, it is common practise to run the raw audio signal through a pre-emphasis filter to compensate the high-frequency part of the speech signal that was suppressed during the human sound production mechanism. To stress the response of a sound we can either amplify dominant frequencies or attenuate components in the surrounding. We do this by applying weights to the according frequencies in the spectrum. In our experiments we tested a two-pole peaking equalizer (EQ) filter, as well as, double-pole lowpass (LP) and highpass (HP) butterworth filters as provided by the sound processing library SoX[1] (Sound eXchange). According filter responses are visualized in Figure 2.

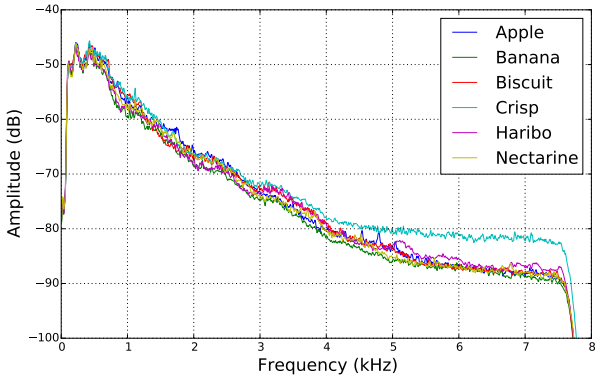

Figure 1: Mean spectrum of the six food classes computed over all files in the training corpus.

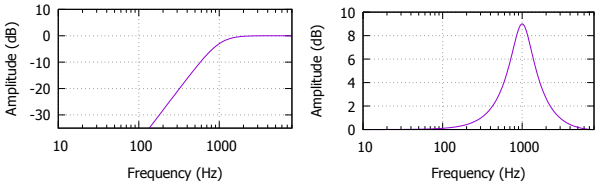

Figure 2: Left: Filter response of a double pole highpass (HP) butterworth filter with a cutoff frequency of 1 kHz. Right: Filter response of a two-pole peaking equalization (EQ) filter with central frequency and band-width set to 1 kHz and a gain of 9 dB.

3.3. Removing Voiced Parts

Though, each file in the corpus is assigned to one food class, it is not said there is a consistent audible impact throughout the recording. Identifying the parts where suspicious sounds such as chewing or smacking are most present may simplify the recognition task. At 2012 Speaker Trait Challenge we have proposed a cluster-based approach to identify frames likely to carry distinctive information [19]. Our experiments showed that dropping non-distinctive frames can lead to an improvement in recognition accuracy. Figure 3 shows spectrograms of the word "endlich" in No_Food and Crisp condition. Comparing the two plots suggests that noise related to food is mainly present in the unvoiced parts. Hence, removing voiced frames may lead to a more compact representation and thereby better recognition rates. To identify voiced frames we extract the fundamental frequency via Sub-Harmonic Sampling (SHS) and apply the Viterbi algorithm to smooth pitch contours and remove octave jumps. We borrow the implementation from OpenSmile (v2.1) [17, 18] and treat frames without a harmonic component as unvoiced (see Figure 4).

4. Results and Discussion

In this section we report and discuss the results we achieved with our proposed approach.

[1] http://sox.sourceforge.net/

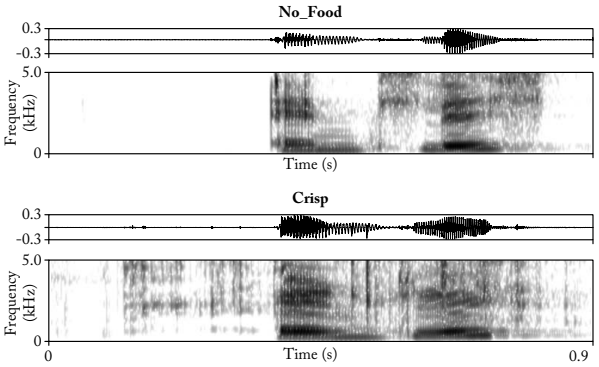

Figure 3: Spectrogram of the word "endlich" pronounced by the first proband in No_Food (top) and Crisp condition (bottom).

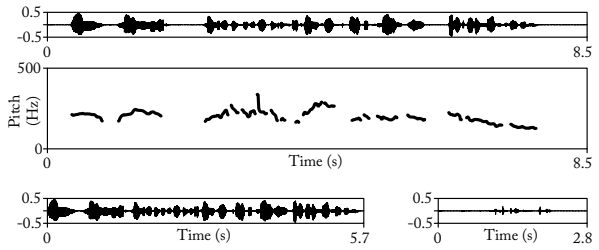

Figure 4: Top: Original waveform. Middle: Fundamental frequency. Bottom: Waveforms when keeping voiced frames (left) or unvoiced frames (right) only.

4.1. Hierarchical Classification

Due to the high number of possible configurations a heuristic is required to pick trees with a good prospect of success. The confusion matrix in Table 1 features No_Food as the by far best recognized class. Splitting off No_Food first and building a separate classifier for the 6 food classes seems promising. The according tree is denoted as **T-L1** and leads to a (marginal) improvement of 0.5%. The class with the second highest recall value is Crisp. Table 1 also tells us it is often confused with Biscuit. It seems natural to group the two classes and split them off next. Tree **T-L2** implements this and achieves an improvement of 1.2%. Among remaining classes, most false predictions occur between Apple and Nectarine. To assign the two classes to a separate branch, tree **T-L3** introduces a third level. It renders a further enhancement of 2.2%. Trees are visualized in Figure 5; according recognition results are found in Table 2.

4.2. Frequency Weighting

Though, the spectrum in Figure 1 suggests that characteristic sounds are primarily located in the high frequencies it is difficult to guess from scratch, which parts of the spectrum should be emphasized to support a certain food type. We therefore tested configurations systematically across the entire spectrum. Following Figure 2 we set bandwidth and gain of the **EQ** filter to 1 kHz and 9 dB, respectively. We then increase the center frequency in steps of 500 Hz and repeat the procedure for **LP** and **HP** filter with different cutoff frequencies. The results are summarized in Figure 6 and confirm our expectation that food sounds are mainly represented in the high frequencies. **EQ** and **HP** filtering especially pays off between 4.0 and 6.5 kHz in the best case yielding an improvement of 4.7%. As expected, chopping off high frequencies with a **LP** filter worsens the results. The best configuration for each filter type is listed in Table 2.

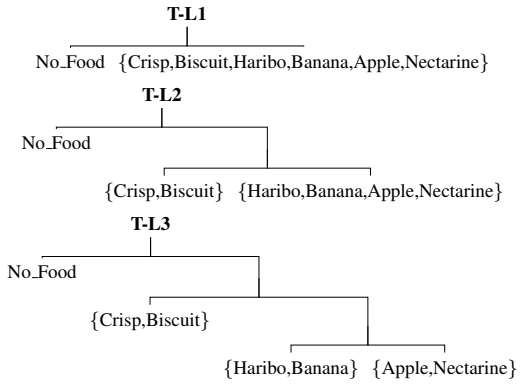

Figure 5: Different tree configurations ordered by tree level.

	NO	AP	HA	NE	BA	CR	BI	Ø
	No_Food	Apple	Haribo	Nectarine	Banana	Crisp	Biscuit	
T-L1	0.0	1.4	-0.8	-0.8	2.1	1.4	0.0	**0.5**
T-L2	0.0	2.1	0.8	1.5	1.4	1.4	0.8	**1.2**
T-L3	0.0	5.0	1.7	2.1	1.4	4.5	0.8	**2.2**
LP{7.0}	2.1	2.9	-3.3	0.8	0.7	2.1	0.8	**0.9**
HP{5.5}	0.7	3.6	5.0	8.3	8.6	2.1	4.5	**4.7***
EQ{0.9}	2.1	6.4	-0.9	0.8	3.6	-1.4	0.0	**1.5**
V	-9.0	-17.9	-22.7	-29.3	-12.1	-10.0	-13.6	**-16.4**
UV	-2.9	-2.9	-8.4	3.8	5.7	-3.6	6.0	**-0.3**
+UV	1.4	7.9	0.0	6.8	4.3	1.4	3.8	**3.7***
T-L3^{EQ,HP}	1.4	12.9	0	15.0	8.6	7.1	8.3	**7.6***
T-L3^{+UV}	0.7	9.3	2.5	9.0	7.1	2.1	5.3	**5.2***
T-L3$^{EQ,HP}_{+UV}$	0.7	10.7	0.9	12.0	3.6	10.0	14.3	**7.5***

Table 2: Recognition rates for tested methods in % relative to baseline in Table 1 (best configuration highlighted in gray). Results yielding a significant improvement according to a McNemar's chi-squared test ($p < 0.05$) [20] are marked with *.

4.3. Removing Voiced Parts

The results in Table 2 support our assumption that food sounds are best recognized from unvoiced regions. Extracting features on unvoiced (**UV**) parts has almost no impact on recognition rates, whereas keeping voiced (**V**) parts only leads to a clear drop of 16.4%. However, it turned out rather surprising that removing pitch related features in both cases caused a drop in performance, even though we would not expect them to carry much useful information if extracted from solely unvoiced frames. Generally, we got the impression that leaving out features impaired the results. On the other hand, when we combined features from both, unvoiced and original files, in a single vector **+UV** (12'746 features) we observed an improvement of 3.7%.

4.4. Combined Approach

Tested methods enhance results by ~2-5%. Can we expect more if we combine them in a single approach? We have seen that recognition rates improve if files are pre-processed with a **EQ** or a **HP** filter, which emphasize the high frequency parts of the spectrum. So far we have evaluated frequency weighting by means of all classes. However, it may work even better when applied on smaller subsets as it allows for a fine-grained adap-

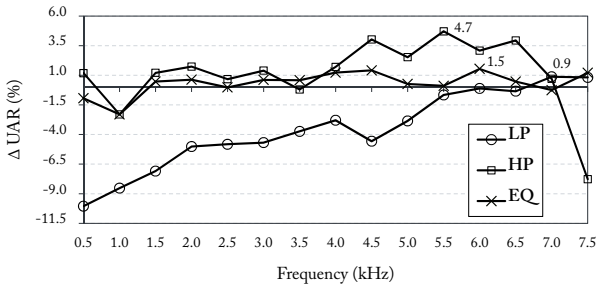

Figure 6: Relative UAR as a function of frequency. In case of **HP** and **LP** cutoff frequency. In case of **EQ** center frequency.

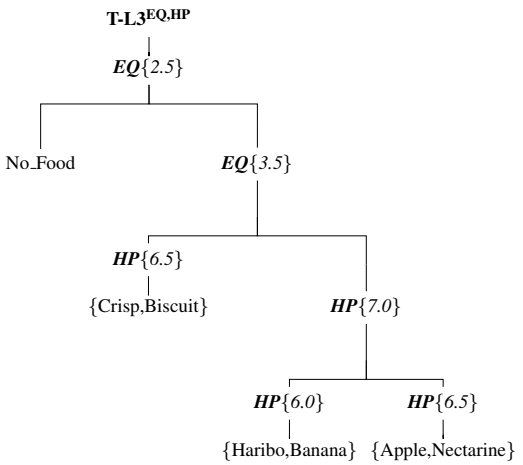

Figure 7: Decomposing a classification problem into a tree structure of independent classification steps offers the possibility to use optimized feature sets at each node. Here, we use **EQ** and **HP** filters to pre-process the input files. Edges are labeled with the configuration that gave best performance in our tests.

tion to the characteristic spectrum of single food classes. And this is where hierarchical classification pays off once again. Namely, because it splits a recognition problem into a series of independent classification steps on increasingly smaller subsets. This allows it to use different **EQ** or **HP** filters at each node and extract a customized feature set. We decided to keep tree **T-L3** as it showed best performance in our experiments and also has the highest number of nodes. To find tailored filter configurations for the nodes we repeated the series of tests in Figure 6 and kept only the samples from classes in the according branch. Figure 7 shows the according tree **T-L3$^{\mathbf{EQ,HP}}$**. The edges of the tree are labeled with the configuration that worked best. At the root node, for instance, an **EQ** filter with a center frequency of 2.5 kHz is installed, whereas for the decision between Apple and Nectarine at the bottom right leaf a **HP** filter with a cutoff frequency of 6.5 kHz is used. Generally, **EQ** filters with a center frequency around 3 kHz were selected at upper tree levels when most of the classes are still included. Towards the leafs with few classes left **HP** filters with cutoff frequencies around 6.5 kHz performed best. With the adapted tree we get an improvement of 7.6%. We also repeated the procedure after removing voiced frames, but in this case filtering appeared to be less useful. We believe this is because in the unvoiced spectrum characteristic food sounds are already emphasized per se. We then tried to enrich the nodes of the original and the adapted tree

	CV Train	**Train/Test**
Baseline	60.5	63.5
+UV	64.2 (+3.7)	64.7 (+1.2)
T-L3	62.7 (+2.2)	66.9 (+3.4)
T-L3$^{\mathbf{EQ,HP}}$	68.1 (+7.6)	67.6 (+4.1)
Challenge Baseline	61.3	65.9

Table 3: Comparison of UAR in % achieved via cross-validation on training set (**CV Train**) and on the test set after using the full training set to build the classifier (**Train/Test**). Relative improvements to the respective baseline are provided in brackets. Official challenge baselines are enclosed in the last row.

with the original **UV** set, but without further measurable success. According trees are denoted as **T-L3$^{\mathbf{+UV}}$** and **T-L3$^{\mathbf{EQ,HP}}_{\mathbf{+UV}}$**, respectively.

4.5. Results on Test Set

Finally, Table 3 reports results on the test set. Since test files are given without ground truth they provide a better hint on the generalisability of the approach. Best results are again achieved with a combined system yielding an improvement of 4.1%. Yet, since the gain is not as high as for the training set, this is a sign that the configurations determined for the training files do not translate perfectly to the test set. A trend that also holds when checking results against the official baseline provided by the challenge organizers [1] (see last row in Table 3).

5. Conclusions

In this paper we presented a new method to detect whether a speaker is eating or not, and if so, which type of food the speaker is currently tasting. The proposed approach exploits that typical sounds like smacking or cracking have a characteristic spectrum and applies frequency-dependent weighting to improve intelligibility of those sounds. It turned out that emphasizing frequencies above 4.0 kHz led to best performances. The approach proved to be especially useful if the recognition problem was decomposed into a hierarchical sequence of binary decisions allowing for a fine-grained adaption to certain food types. We also figured that meaningful information was mainly found in the unvoiced parts of the signals. With the proposed approach we were able to improve results in a 7-class recognition problem (6 types of food and no food) by more than 7% on the training set (using leave-one-eater-out cross validation) and 4% on the test set, respectively. In the study at hand we experimented with few basic filters to either amplify or attenuate certain frequency bands. Due to the promising results it might be worthwhile to apply a more complex equalization. However, this requires some intelligent method to determine the optimal filter response (for instance via correlation analysis) since it is not feasible to test the vast number of possible configurations in a brute force way.

6. Acknowledgements

The work described in this paper has received funding from the European Commission under the contract number H2020-RIA-645012, KRISTINA, and the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378, ARIA-VALUSPA.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinsons & eating condition," in *to appear*, 2015.

[2] S. Hantke, F. Weninger, R. Kurle, A. Batliner, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food type," in *to appear*, 2015.

[3] B. Schuller, M. Wöllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement," *EURASIP J. Audio Speech Music Process.*, vol. 2009, pp. 5:1–5:17, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/942617

[4] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261 – 291, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939400059J

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, vol. 4, Apr 1979, pp. 208–211.

[7] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, no. 1-3, pp. 133–147, Aug. 1998. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(98)00033-8

[8] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer Speech & Language*, vol. 21, no. 3, pp. 443 – 457, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230806000301

[9] M. Seltzer, A. Acero, and K. Kalgaonkar, "Acoustic model adaptation via linear spline interpolation for robust speech recognition," in *ICASSP*. IEEE, March 2010. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=130977

[10] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572 – 585, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639313000046

[11] C. Dacremont, "Spectral composition of eating sounds generated by crispy, crunchy and crackly foods," *Journal of Texture Studies*, vol. 26, no. 1, pp. 27–43, 1995. [Online]. Available: http://dx.doi.org/10.1111/j.1745-4603.1995.tb00782.x

[12] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. André, "Age and gender classification from speech using decision level fusion and ensemble based techniques." in *INTERSPEECH*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010, pp. 2798–2801. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2010.html#LingenfelserWVKA10

[13] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, "Analysis of chewing sounds for dietary monitoring," in *Proceedings of the 7th International Conference on Ubiquitous Computing*, ser. Lecture Notes in Computer Science, vol. 3660. Springer-Verlag, 2005, pp. 56–72.

[14] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "Bodybeat: A mobile system for sensing non-speech body sounds," in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '14. New York, NY, USA: ACM, 2014, pp. 2–13. [Online]. Available: http://doi.acm.org/10.1145/2594368.2594386

[15] B. W. Schuller, "The computational paralinguistics challenge [social sciences]," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 97–101, 2012. [Online]. Available: http://dx.doi.org/10.1109/MSP.2012.2192211

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] 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*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874246

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502224

[19] J. Wagner, F. Lingenfelser, and E. André, "A frame pruning approach for paralinguistic recognition tasks." in *INTERSPEECH*. ISCA, 2012.

[20] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947. [Online]. Available: http://dx.doi.org/10.1007/BF02295996