"You sound ill, take the day off": Automatic Recognition of Speech Affected by Upper Respiratory Tract Infection

Nicholas Cummins[1], Maximilian Schmitt[1], Shahin Amiriparian[1,2], Jarek Krajewski[4,5], Björn Schuller[1,3]

***Abstract*— A combination of passive, non-invasive and non-intrusive smart monitoring technologies is currently transforming healthcare. These technologies will soon be able to provide immediate health related feedback for a range of illnesses and conditions. Such tools would be game changing for serious public health concerns, such as seasonal cold and flu, for which early diagnosis and social isolation play a key role in reducing the spread. In this regard, this paper explores, for the first times, the automated classification of individuals with Upper Respiratory Tract Infections (URTI) using recorded speech samples. Key results presented indicate that our classifiers can achieve similar results to those seen in related health-based detection tasks indicating the promise of using computational paralinguistic analysis for the detection of URTI related illnesses.**

***Index Terms*— Upper Respiratory Tract Infection, Classification, Paralinguistic Analysis, Bag-of-Audio-Words, Feature Selection**

I. INTRODUCTION

Upper Respiratory Tract Infections (URTIs) are a serious public health concern. URTIs are caused by a range of illnesses such as the Common Cold and Influenza (flu), both of which spread easily through a population. The World Health Organisation estimates that cold and flu epidemics result in approximately 3 to 5 million cases of severe illness, and about 250 000 to 500 000 thousand deaths per year [1]. Technology can play a key role in helping prevent the spread of illnesses related to URTIs through early detection. For example, using social media platforms such as Twitter, it is possible to track the spread of seasonal influenza epidemics in a population and provide timely warnings to public health authorities and at risk individuals [2].

One technology yet to be realised as an early diagnosis system for potential URTIs is automated speech analysis. Speech, as a complex and highly sensitive output system, is potentially well suited for the remote diagnosis of URTI related conditions. Slight changes in a speaker's physical and mental state are known to affect the muscular systems used to control vocal apparatus altering the acoustic properties of the resulting speech. Speech analysis paradigms, nominally based on high dimensional paralinguistic feature representations in combination with machine learning, has matured into a new form of active and passive remote sensing technology suitable for a broad range of health conditions [3].

This paper explores, if state-of-the-art paralinguistic analysis paradigms can be used to classify speech affected by an URTI. To the best of the authors' knowledge, this is the first time such a study has been undertaken. We test the suitability of two such paradigms; a brute-force based system which utilises the widely used COMPARE feature set, introduced at Interspeech 2013 *Computational Paralinguistics Challenge* [4], in combination with *Support Vector Machines* (SVMs), and, a state-of-the-art *Bag-of-Audio-Words* (BoAW) based system [5]. Further, we also explore the advantages of performing feature selection in both paradigms.

The rest of this paper is laid out as follows: Section II gives a brief overview of related works, and Section III introduces the Upper Respiratory Tract Infection Corpus. Our proposed methods are outlined in Section IV, and the experimental results and subsequent discussion are presented in Section V and Section VI respectively. Finally, a brief conclusion and outline of our future work plans are given in Section VII.

II. RELATION TO PRIOR WORK

The combination of the COMPARE feature set and a SVM can be considered as a default-standard system in computational paralinguistics; as evidenced by its use in the popular Interspeech Computational Paralinguistics Challenges. Whilst this set-up has not been tested for its efficacy in recognising speech affected by a URTI, it has been used in other similar speech-based health recognition tasks including Autism detection [4], cognitive and physical load classification [6], and Parkinson's Condition detection [7].

BoAW is starting to gain popularity in many computational paralinguistics tasks. For example, BoAW has been used in related tasks including snore sound classification to aid the detection of Obstructive Sleep Apnoea [8]. It has also been shown to be useful in speech-based autism detection [9], or depression detection [10]. Finally, it has produced state-of-the-art results when performing emotion detection [11].

III. URTI DATASET

The *Upper Respiratory Tract Infection Corpus* (URTIC) was created at the Institute of Safety Technology, University of Wuppertal, Germany. The corpus consists of speech from 630 participants (382 m, 248 f), with a total of 11 283 audio

[1]Nicholas Cummins, Maximilian Schmittt, Shahin Amiriparian and Björn Schuller are with the Chair of Complex & Intelligent Systems, Univeristy of Passau, Passau, Germany {nicholas.cummins, maximilian.schmitt, shahin.amiriparian, bjoern.schuller}@uni-passau.de

[2]Shahin Amiriparian is also with the Machine Intelligence & Signal Processing Group, Technische Universität München, Germany

[3]Björn Schuller is also with the Machine Learning Group, Imperial College London, U.K.

[4]Jarek Krajewski is with the Institute of Safety Technology, Human Factors and Diagnostics, University of Wuppertal, Wuppertal, Germany krajewsk@uni-wuppertal.de

[5]Jarek Krajewski is also with Engineering Psychology, Rheinische Fachhochschule Cologne, Cologne, Germany

TABLE I

UPPER RESPIRATORY TRACT INFECTION CORPUS (URTIC): NUMBER OF SUBJECTS PER CLASS IN THE TRAIN/DEVELOPMENT(DEVEL) SPLIT; C: COLD; NC: NON-COLD; F: FEMALE; M: MALE. AT THE TIME OF WRITING THE MAKE UP OF THE TEST PARTITION WAS NOT PUBLICLY AVAILABLE.

#	Train		Devel		Σ
	F	M	F	M	
C	17	20	16	21	74
NC	65	108	66	107	346
Σ	82	128	82	128	420
Σ	210		210		420

TABLE II

UPPER RESPIRATORY TRACT INFECTION CORPUS (URTIC): NUMBER OF INSTANCES PER CLASS IN THE TRAIN/DEVELOPMENT(DEVEL) SPLITS; C: COLD; NC: NON-COLD; F: FEMALE; M: MALE. AT THE TIME OF WRITING THE TEST LABELS WERE NOT PUBLICLY AVAILABLE.

#	Train		Devel		Σ
	F	M	F	M	
C	389	581	460	551	1 981
NC	2 821	5 714	2 978	5 607	17 120
Σ	3 210	6 295	3 438	6 158	19 101
Σ	9 505		9 596		19 101

recordings. The mean age of the participants was 29.5 years, with a standard deviation of 12.1 years and a range of 12 to 84 years.

All recordings were made in quiet rooms with a microphone/headset/hardware setup, the tasks were presented on a computer in front of the participants. Audio files were recorded with a 44.1 kHz sample rate and, down-sampled to 16 kHz with a quantization of 16 bit. The speech material consisted of different reading passages and speaking tasks. The participants were asked to read aloud sentences regarding voice commands as used for driver assistance systems and short stories like "The North Wind and the Sun" (widely used within phonetics), and "Die Buttergeschichte" (standard reading passage in German, used in speech/language pathology). Furthermore, spontaneous narrative speech was elicited by asking subjects to briefly comment on, e. g., their last weekend, the best present they ever received or to describe a picture. Each session lasted between 15 minutes to 2 hours.

Each participant had to report a binary one-item measure of having a cold on the German version of the *Wisconsin Upper Respiratory Symptom Survey* (WURSS-24) [12]. The questionnaire is an evaluative illness-specific quality of life instrument and assesses the symptoms of the common cold. In order to investigate cold induced speech changes, the primary outcome of interest was the global illness severity item (on a scale of 0 = not sick to 7 = severely sick).

The corpus consists of approximately 45 minutes of speech and the available recordings were split into 28 652 chunks. According to the binary one-item measures, the chunks are split into two classes; chunks with a corresponding WURSS-24 equal to zero were assigned to *Non-Cold* (NC), whilst chunks with a corresponding WURSS-24 greater than zero were assigned to *Cold* (C). The chunks are then sub-divided (in a speaker independent manner) into *Train*, *Development*, and *Test* partitions. The division of the participants and their gender between the Train and Development partitions is given in Table I, and the distributions of the chunks between the Train and Development partitions is also provided in Table II[1].

TABLE III

THE 65 LOW-LEVEL DESCRIPTORS (LLD) PROVIDED IN THE COMPARE ACOUSTIC FEATURE SET.

4 energy related LLD	**Group**
Sum of Auditory Spectrum (Loudness)	prosodic
Sum of RASTA-filtered Auditory Spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filtered Auditory Spectral Bands 1–26 (0-8 kHz)	spectral
MFCC 1–14	cepstral
Spectral Energy 250–650 Hz, 1 kHz–4 kHz	spectral
Spectral Roll-Off Point 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F0 (SHS & Viterbi Smoothing)	prosodic
Probability of Voicing	voice quality
log. HNR, Jitter (local & DDP), Shimmer (local)	voice quality

IV. EXPERIMENTAL SETTINGS

A. ComParE Acoustic Feature Set

All results presented are based on the *Interspeech 2013 Computational Paralinguistics Challenge* feature set COMPARE. This feature set contains 6 373 static features (i. e., functionals) of *low-level descriptor* (LLD) contours. An overview of the prosodic, spectral, cepstal, and voice quality LLDs is given in Table III. The functionals applied to the LLD contours include the mean, standard deviation, percentiles and quartiles, linear regression functionals, and local minima/maxima related functionals; for full details the reader is referred to [13].

B. Bag-of-Audio-Words

Bag-of-Audio-Words (BoAW) is based on the widely used bag-of-words approach from natural language processing, where documents are classified based on a histogram representation of linguistic features. BoAW involves quantisation of acoustic LLDs (cf. Figure 2), where each frame-level LLD vector is assigned to an audio word from a previously learnt codebook. Counting the number of assignments for each audio word, a fixed length histogram (bag) representation of an audio clip is generated. For codebook generation, a random sampling of a certain number of LLDs from

[1]At the time of writing the URTIC was an active Interspeech Computational Paralinguistics Challenge dataset, therefore the equivalent division for the test partitions were not publicly available

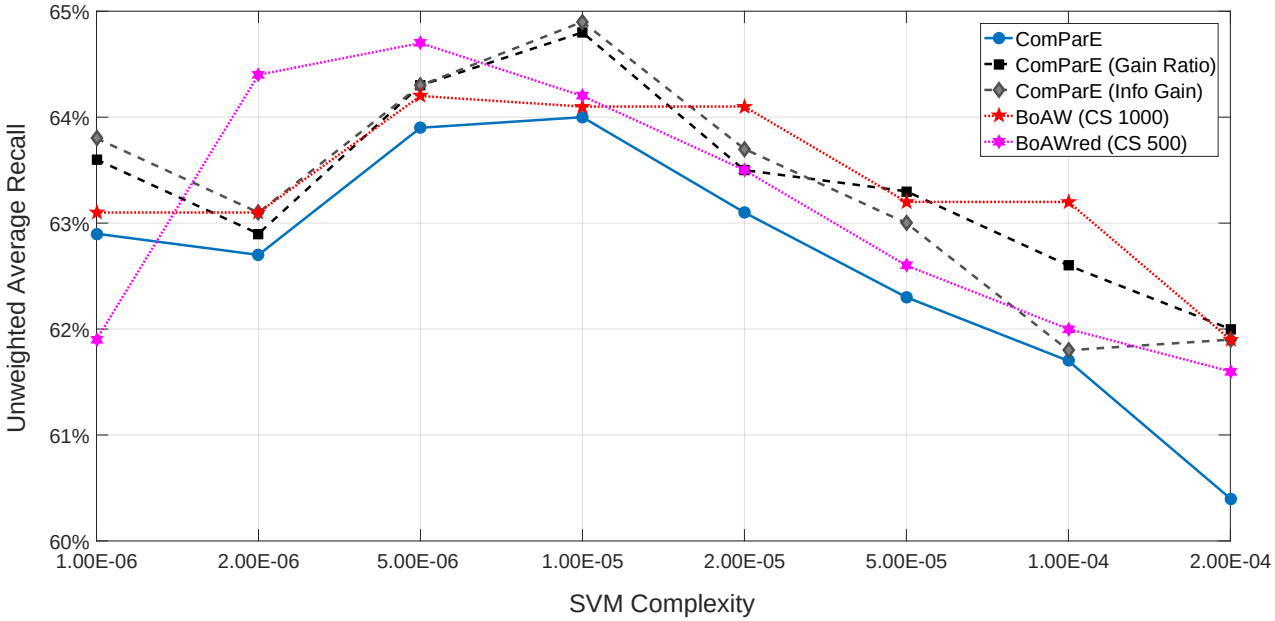

Fig. 1. A Comparison of UARs found on the URTIC Corpus development partition over a range of different SVM complexities. The feature representation are: ComParE, ComParE with Information Gain Ratio Feature Selection (Gain Ratio), ComParE with Information Gain Feature Selection (Info Gain), a ComParE–LLD Bag-of-Audio-Words (BoAW) with a codebook size of 1000 ($Cs = 1000$), and a reduced set of ComParE–LLDs BoAWred ($Cs = 500$).

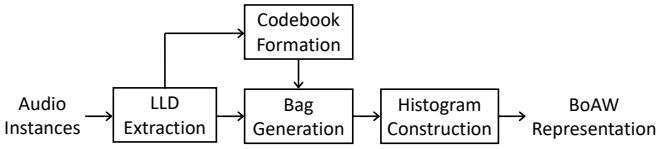

Fig. 2. A generalised overview of the bag-of-audio-words formation.

the training data has proven to be suitable [11]. Due to the quantisation step, BoAW representations are generally considered more robust than LLDs. The BoAW features have been computed using the toolkit openXBOW [5], which provides crossmodal (acoustic, visual, and text) bag-of-words generation.

All BoAW representations were generated from the 65 LLDs from the ComParE feature set with the corresponding deltas. For each of the LLDs and their deltas, a separate codebook was learnt using random sampling of the LLDs from the training data. An extensive iterative search of *codebook size* (*Cs*), between 250 and 8000 audio words, was also conducted (results not given). In order to get rid of the variation of scales between LLDs, which might have an influence on the quantisation step, LLDs were normalised to zero mean and unit variance. The parameters mean and standard deviation have been estimated from the training data (online approach).

C. Classification Set-Up

All results are reported in terms of *Unweighted Average Recall* (UAR); this is the standard measure of the Interspeech Computational Paralinguistics Challenges and is suitable for use when the distribution among classes is not balanced. In the training data, the instances of the minority class (Cold) were upsampled by a factor of 9 to overcome potential effects of class imbalance.

All feature vectors or BoAW representations were fed into an SVM classifier with a linear kernel (SMO implementation in WEKA [14], with standardisation). The complexity parameter (C) was optimised on a scale from $2 \cdot 10^{-6}$ to $2 \cdot 10^{-4}$ based on the on the development data. Note, initial experiments on a larger range of C values (results not given) indicated optimal performances were in this smaller C range. For the final evaluation on the test data, a model was trained on the fused training and development data.

Given the high dimensionality of the ComParE feature set, we also tested the following two different feature reduction methods. Namely, *Information Gain Ratio Feature Selection* (Gain Ratio) and *Information Gain Feature Selection* (Info Gain), both available in the WEKA toolkit [14]).

V. Results

As can be seen in Figure 1, all paradigms tested achieved very similar performances on the development set across the different SVM complexity parameters. The ComParE features' highest UAR, 64.0%, was found with $C = 1.00E-05$. Small gains in performance were achieved when applying feature selection. Both methods achieved almost identical performance; Gain Ratio achieved a UAR of 64.8% ($C = 1.00E-05$) and Info Gain a UAR of 64.9% ($C = 1.00E-05$). For the BoAW representation, our initial testing (results not given) revealed that $Cs = 1000$ gave the strongest and most consistent results across the different SVM complexities. The highest BoAW UAR achieved was 64.2% ($C = 1.00E-06$); as can be seen in Figure 1, this is slightly higher than the one achieved by ComParE features but below the maximum UARs of the two feature selection paradigms.

Given the small improvement gained by using feature selection with ComParE, we also tested feature selection

TABLE IV
A COMPARISON OF UARS FOUND ON THE URTIC CORPUS TEST PARTITION FOR DIFFERENT COMPARE AND BOAW BASED FEATURE REPRESENTATIONS.

Feature Representation				
COMPARE	COMPARE (Gain Ratio)	COMPARE (Info Gain)	BOAW (Cs = 1000)	BOAW (Reduced)
70.2%	69.4%	69.3%	67.3%	**70.2%**

with BoAW (BoAWred in Figure 1). Reducing irrelevant information from the input space makes sense for BoAW, as each LLD has the same weight in the quantisation step. Inspecting both sets of COMPARE features selected, we observed that the 6 voicing related LLDs (cf. Table III) were never chosen. Therefore, we split the input space into the 4 groups (with their corresponding deltas); *energy-related*, *RASTA*, *MFCC*, and further *spectral* LLDs. Learning a different codebook for each group ($Cs = 100$ for energy-related, $Cs = 500$ for the other groups) increased the best performing BoAW UAR to 64.8% ($C = 5.00E-06$).

The test set results for the five different paradigms are given in Table IV, the corresponding C values were those identified on the development set. Interestingly, the COMPARE features performed the strongest, although these results are almost identical to those found with the two feature selection methods. BoAW performed the weakest; however, the the combination of BoAW and the reduced LLD set matched performance with COMPARE. A finer search of the complexity space for the Gain Ratio and Info Gain revealed slight improvements in performance could be found, both achieving maximum UARs of 70.4% ($C = 4E-05$).

VI. DISCUSSION

As, to the best of the authors' knowledge, this is the first time that speech based classification of speech affected by a URTI has been performed, we cannot compare the performance of our system with other results in the literature. However, the combination of COMPARE and a SVM has been used in other 2-class speech-based health classification tasks. For example, when used to detect Autism, this paradigm yielded a UAR of 67.1% [4], whilst UARs of 61.6% and 71.9% were achieved for cognitive and physical load respectively [6]. Similarly, BoAW have achieved a UAR of 79.5% for snore sound classification [8]. The closeness in performance to these other (more established) speech based detection tasks indicates the potential for using speech as marker of URTIs.

VII. CONCLUSION

Smart monitoring technologies can play a key role in helping to prevent the spread of commonly occurring diseases, such as the Common Cold and Influenza, by providing simple to use early detection systems, such as a mobile app. The highest observed UAR of 70.2% on the test set of the newly gathered *Upper Respiratory Tract Infection Corpus* indicates the potential of using speech as such a marker.

Future work will include the use of different feature representations as well as more sophisticated classification techniques and features [15]. Given the slight improvement seen when performing feature selection, we will repeat the analysis with specific LLD feature groupings to gain insights into which specific acoustic or prosodic features capture the more salient properties of speech affected by a URTI.

VIII. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 645094 (IA SEWA).

REFERENCES

[1] "World Health Organization (WHO)," http://www.who.int/mediacentre/factsheets/fs211/en/, accessed: 28-01-2017.

[2] H. Achrekar, A. Gandhe, R. Lazarus *et al.*, "Predicting flu trends using twitter data," in *2011 IEEE Conf. on Comp. Comm. Workshops (INFOCOM WKSHPS)*. Shanghai, P. R. China: IEEE, 2011, pp. 702–707.

[3] B. Schuller, T. H. Falk, V. Parsa, and E. Nöth, "Introduction to the Special Issue on Atypical Speech & Voices: Corpora, Classification, Coaching & Conversion," *EURASIP Jour. on Audio, Speech, and Music Proc., Special Issue on Atypical Speech & Voices: Corpora, Classification, Coaching & Conversion*, vol. 29, pp. 100–131, 2015.

[4] B. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[5] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *arxiv.org*, no. 1605.06778, May 2016, 9 pages.

[6] B. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *Proc. of INTERSPEECH*. Singapore: ISCA, September 2014.

[7] ——, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition," in *Proc. of INTERSPEECH*. Dresden, Germany: ISCA, September 2015, pp. 478–482.

[8] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds excitation localisation," in *Proc. 14th ITG Conference on Speech Comm.*, ser. ITG-Fachbericht, vol. 267. Paderborn, Germany: IEEE/VDE, October 2016, pp. 264–268.

[9] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards Cross-lingual Automatic Diagnosis of Autism Spectrum Condition in Children's Voices," in *Proc. 14th ITG Conference on Speech Comm.*, ser. ITG-Fachbericht, vol. 267. Paderborn, Germany: IEEE/VDE, October 2016, pp. 264–268.

[10] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Jour. on Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[11] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. of INTERSPEECH*. San Francisco, CA, USA: ISCA, 2016, pp. 495–499.

[12] B. Barrett, K. Locken, R. Maberry *et al.*, "The Wisconsin Upper Respiratory Symptom Survey (WURSS): a new research instrument for assessing the common cold," *The Jour. of family practice*, vol. 51, no. 3, p. 265, March 2002.

[13] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[15] S. Amiriparian, J. Pohjalainen, E. Marchi, S. Pugachevskiy, and B. Schuller, "Is deception emotional? an emotion-driven predictive approach," in *Proc. of INTERSPEECH*. San Francisco, CA, USA: ISCA, 2016, pp. 2011–2015.