DETERMINING THE BEST ACOUSTIC FEATURES FOR SMOKER IDENTIFICATION

Zhizhong Ma[1], Yuanhang Qiu[1], Feng Hou[1], Ruili Wang[1], Joanna Ting Wai Chu[2], Christopher Bullen[2]*

[1]School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand
[2]National Institute for Health Innovation, University of Auckland, Auckland, New Zealand

ABSTRACT

Speech-based automatic smoker identification (also known as smoker/non-smoker classification) aims to identify speakers' smoking status from their speech. In the COVID-19 pandemic, speech-based automatic smoker identification approaches have received more attention in smoking cessation research due to low cost and contactless sample collection. This study focuses on determining the best acoustic features for smoker identification. In this paper, we investigate the performance of four acoustic feature sets/representations extracted using three feature extraction/learning approaches: (i) hand-crafted feature sets including the extended Geneva Minimalistic Acoustic Parameter Set and the Computational Paralinguistics Challenge Set, (ii) the Bag-of-Audio-Words representations, (iii) the neural representations extracted from raw waveform signals by SincNet. Experimental results show that: (i) SincNet feature representations are the most effective for smoker identification and outperform the MFCC baseline features by 16% in absolute accuracy; (ii) the performance of hand-crafted feature sets and the Bag-of-Audio-Words representations rely on the scale of the dimensions of feature vectors.

Index Terms— Smoker identification, acoustic features, ComParE, BoAW, SincNet

1. INTRODUCTION

Speech-based automatic smoker identification (also known as smoker/non-smoker classification) aims to identify a speaker's smoking status from his or her speech data. Automatic smoker identification has a variety of applications including smoking status validation [1], smoking cessation tracking [2] and speaker profiling [3]. Speech-based smoker identification has advantages over traditional biochemical measures for determining if an individual has successfully stopped smoking, because of the costs and the ease of the sample collection process. Speech-based automatic smoker identification is especially useful for smoking cessation research under the current COVID-19 pandemic where movement restrictions may make other methods more difficult or expensive than usual.

*Corresponding author

Many studies have shown that cigarette smoking negatively affects smokers' vocal tissues and permanently alters the acoustic properties of smokers' speech compared with non-smokers [4]–[7]. Such alterations are confirmed by assessing the acoustic features like fundamental frequency (F_0), jitter and shimmer [8]–[11]. There has been some previous work in speech-based automatic smoker identification field [3], [12]. In these two papers, the authors utilised Mel Frequency Cepstral Coefficients (MFCC) to identify smokers from their spontaneous speech. Recently, in the speech-health analysis related field, hand-crafted feature sets and learned neural representations, which were not considered for smoker identification, have been proven to be more effective acoustic features than MFCC [13]–[16].

The hand-crafted feature sets (e.g., eGeMAPS [17] and ComParE [18]) and the Bag-of-Audio-Words (BoAW) [19] representations have been used successfully for speech-health analysis related tasks [13], [14], [16], [20]. Furthermore, learning task-driven features directly from the raw waveform by deep neural networks (DNNs) has been proven to be an effective feature extractor [21] for a variety of applications, such as speech recognition [22], speaker recognition [23] and emotion recognition [24]. For example, SincNet [23] is a Convolutional Neural Network (CNN) for learning feature representations from raw waveforms. Compared with hand-crafted feature sets, SincNet is more effective in learning the most suitable feature representations for the given tasks [15], [23].

We hypothesise that the quality of the acoustic features is crucial for the performance of speech-based smoker identification systems. In this study, we aim to identify speakers' smoking status by using more advanced feature extraction/learning approaches. We compare the four acoustic feature sets/representations extracted/learned using three feature extraction/learning approaches: (i) hand-crafted feature sets, i.e., eGeMAPS and ComParE, (ii) the BoAW representations quantising acoustic low-level descriptors (LLDs), (iii) the neural representations extracted from raw waveform signals by SincNet.

However, there are just a few publicly available datasets for smoker identification tasks. The dataset we utilise is derived from the two corpora (i.e., the Mixer 4 and 5 Speech Corpus [25], and the Mixer 6 Speech Corpus [26]) which include rich metadata regarding speakers' smoker status, age,

height, weight, etc., making them applicable for smoker identification experiments.

The main contributions of our paper are as follows:

• We identify that the most effective acoustic features are the feature representations learned by using deep neural networks.

• We compare the effectiveness and generalizability of acoustic features extracted using three different feature extraction/learning approaches for smoker identification.

• We propose a new dataset for smoker identification experiments based on two existing corpora.

The rest of this paper is structured as follows. Section 2 presents the related work of the acoustic feature sets/representations we utilise in this paper. Section 3 presents our dataset. The design and methods are provided in Section 4. Section 5 describes the experimental results, and the conclusions and proposals for future work are discussed in Section 6.

2. RELATED WORK

2.1. extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)

eGeMAPS is a low-dimensional, frame-level, knowledge-inspired acoustic feature set containing a wide range of standardised relevant acoustic features [17]. eGeMAPS is extracted on two levels: (i) low-level descriptors (LLDs), (ii) statistical functionals.

eGeMAPS includes 88 acoustic features derived from 23 LLDs that covers spectral, cepstral, prosodic and voice quality information of the speech, as shown in Table 1. The efficiency of eGeMAPS has been proven successful in various areas of clinical and paralinguistic speech analysis, including Alzheimer's Dementia detection [13], speech intelligibility assessment [14] and speech emotion recognition [27].

2.2. Computational Paralinguistics Challenge set (ComParE)

ComParE is a well-evolved, high-dimensional brute-forced acoustic feature set that is extracted on three levels: (i) low-level descriptors (LLDs), (ii) statistical functionals, and (iii) LLDs deltas [18]. It contains 6373 static features resulting from the computation of various functionals over 65 LLDs. ComParE consists of fundamental frequency (F_0), energy, spectral, cepstral coefficients (MFCCs) and voicing related frame-level features. It also includes zero-crossing rate, jitter, shimmer, harmonic-to-noise ratio (HNR), spectral harmonicity and psychoacoustic spectral sharpness. The statistical functionals applied to the LLDs include the mean, standard deviation, percentiles and quartiles, linear regression functionals, and local minima/maxima related functionals.

The ComParE feature set has demonstrated its ability and robustness for capturing acoustic information in many speech-health analysis related tasks, including COVID-19 diagnosis [16] and upper respiratory tract infections (URTI) classification [28].

Table 1. Summary of feature sets/representations utilised in this study

Name	Type	No. of Features
eGeMAPS	Hand-crafted	88
ComParE	Hand-crafted	6373
BoAW	Hand-crafted+BoAW	1000
SincNet	Raw Waveform	2048

2.3. Bag-of-Audio-Words (BoAW)

BoAW is extended from the concept of Bag-of-Words [29], a common representation of information in the Natural Language Processing (NLP) field. BoAW is a sparse audio representation that first clusters the input frame-level feature vectors (e.g., MFCC, eGeMAPS, ComParE), replaces each frame-level feature vector by its cluster, and then uses a rich dictionary (i.e. codebook) of these clusters to represent an utterance-level feature vector [19]. The main advantage of BoAW is its capacity of summarising the meaningful information of a variable-length input audio using a fixed-length vector (i.e., the histogram). The histogram represents the distribution of quantised feature vectors from a given audio instance [30].

Recently, the BoAW representation approach has become very popular and has demonstrated its suitability in various speech related fields [20], [30], [31].

2.4. SincNet

SincNet is a novel CNN-based architecture, originally proposed for speaker recognition [23]. SincNet has embedded bandpass filters for extracting features from the raw waveform. Instead of learning all elements from each filter in the traditional CNN architecture, SincNet only learns those low and high cutoff frequencies from raw waveform, which makes it more interpretable and faster to converge.

SincNet has shown improved performance for research in different areas of speech-related tasks, including neurodegenerative related disorder classification [15], speech-based age and cognitive decline estimation [32] and speech emotion recognition [33].

3. DATASET

In the absence of large-scale, well-designed datasets expressly for smoker identification experiments, we collect and create our datasets by extracting from two corpora released through the Linguistic Data Consortium (LDC): (i) the Mixer 4 and 5 Speech Corpus, (ii) the Mixer 6 Speech Corpus. Both corpora comprise conversation recordings made via the public telephone network and multiple microphones in office-room settings. The main difference in the setting is that few of the 616 distinct speakers in the Mixer 4 and 5 Speech

Corpus are bilingual English speakers, while the rest of the speakers of Mixer 4 and 5 Speech Corpus and all 594 distinct speakers in the Mixer 6 Speech Corpus are native English speakers. There is no overlap between the two corpora.

However, not all speakers in these two corpora have a valid smoking status label. In Mixer 4 and 5 Speech Corpus, only 89 of 616 speakers have smoking status labels. There are 40 female smokers, 8 female non-smokers, 37 male smokers, 4 male non-smokers. In Mixer 6 Speech Corpus, 589 of 594 speakers have smoking status labels. There are 48 female smokers, 252 female non-smokers, 70 male smokers, 219 male non-smokers. For valid smoker identification purposes and balancing speakers' gender and smoker status distribution, only those speakers with valid smoking status labels are considered in our experiments. In the end, 200 speakers (50 female smokers, 50 female non-smokers, 50 male smokers, 50 male non-smokers) are selected for experiments. The details of the speaker's status are shown in Table 2 below. Most of the speakers have two or more 12 mins to 15 mins transcripts reading audio segments; a few only have one 12 mins transcript reading audio segment. We split the training set, development set and test set following the 8:1:1 ratio. We chose 5 female smokers, 5 female non-smokers, 5 male smokers, 5 male non-smokers as the test set and the rest speakers for the training set and the development set. To ensure our smoker identification experiments are speaker-independent, recordings from speakers who contributed more than one recording are retained in the same division.

Table 2. The status of the speaker's age in our dataset.

	Avg	Min	Max
Female Smokers	31.76	18	63
Female Non-Smokers	31.36	17	68
Male Smokers	30.38	19	60
Male Non-Smokers	28.10	19	60

4. METHODS AND EXPERIMENTS

4.1. Feature Extraction

Before extracting any acoustic features, we normalise the volume of all voice utterances into the range [-1: +1] dBFS. The goal is to improve the smoker identification's robustness against diverse recording conditions, such as microphone distance from the subject's mouth.

We use the openSMILE toolkit [18] and standard configuration files (i.e., eGeMAPSv01a.conf, ComParE_2016.conf) to extract features for eGeMAPS and ComParE standard feature sets, respectively. We also use an MFCC feature set (MFCC12_0_D_A.conf) as our baseline feature set. The openXBOW toolkit [19] is used to generate BoAW representations from the 23 LLDs of eGeMAPS and the 65 LLDs of ComParE with the corresponding deltas, respectively. For each of the LLDs and their deltas, a separate codebook is learnt using random sampling of the LLDs from the training data. We test codebook sizes of N =500, 1000 and 5000. In order to get rid of the variation of scales between LLDs, which have an influence on the quantisation step, LLDs are normalised to zero mean and unit variance. The parameters mean and standard deviation have been estimated from the training. We observe that the ComParE-BoAW feature representations with a codebook size of 1000 achieved the best performance, hence it is reported in the rest of our experiments.

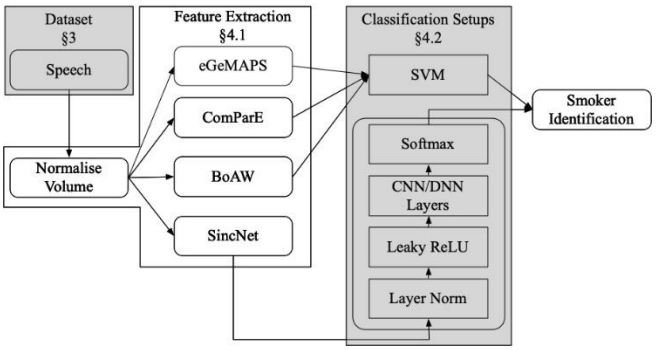

Fig.1. Our proposed methodology.

The SincNet layer is applied to the raw waveform and acts as a feature extractor to generate feature vectors. The raw waveform of each speech recording is chunked using a frame size of 200 ms and feed into the SincNet architecture described in Section 4.2.

4.2. Classification Setups

For evaluating the extracted hand-crafted feature sets, Support Vector Machine (SVM) is utilised because of its high effectiveness in the acoustic-based speech classification fields [13], [28], [34]. For SVM, we set the cost parameter C as 0.01 and use Radial Basis Function (RBF) kernels. The SVM classifier is trained by using the hand-crafted feature sets extracted from the training and development sets. The test set is used for evaluating the performance of the SVM classifier. SVM is implemented with scikit-learn[1].

As illustrated in Fig.1, all neural representations extracted by SincNet are fed into a CNN classifier. The CNN classifier we implemented was proposed by Ravanelli and Benjio [23], which has two standard convolutional layers, each with 60 filters of length 5 to evaluate the neural representations. For both the input samples and all convolutional layers (including the SincNet input layer), layer normalisation [35] is employed. Following that, three fully-connected layers with a total of 2048 neurons are applied and normalised with batch normalisation [36]. Leaky-ReLU [37] (with variable nonlinearity) have been used in all hidden layers. The neural networks are implemented with PyTorch[2].

[1]https://scikit-learn.org
[2]https://pytorch.org

5. RESULTS AND DISCUSSION

A summary of experimental results for smoking identification is provided in Table 3. Our results show that all four proposed acoustic feature sets/representations achieve better performances on the dataset than the MFCC baseline features.

Table 3. Experimental results of different acoustic feature sets/representations on the test set

Features	Accuracy	F1-score
MFCC	0.71	0.70
eGeMAPS	0.78	0.77
ComParE	0.83	0.83
BoAW*	0.81	0.80
SincNet	**0.87**	**0.87**

For the hand-crafted feature sets, the ComParE feature set (6373 features) achieves the higher classification accuracy with 83%, which is significantly better than the MFCC baseline of 71%. The eGeMAPS feature set (88 features) achieves a classification accuracy of 78%. The performance is higher than the MFCC baseline feature but is slightly lower than the one achieved by ComParE. This indicates that hand-crafted feature sets including fundamental frequency (F_0), jitter, shimmer etc., provide better performance than traditional conventional acoustic features such as MFCC in this task-driven speech classification experiment. It also suggests that the more features in the hand-crafted feature sets are used, the better the classification performance will be.

The BoAW representation approach (i.e., ComParE-BoAW) achieves a slightly lower performance with an accuracy of 81% compared with using the ComParE feature set directly. We also test the performance of BoAW built from the eGeMAPS feature set, but the results are consistently lower than using the eGeMAPS feature set directly and are not included in Table 3. A key direction for future research is determining the most useful frame-level features for a BoAW model.

Compared with hand-crafted features, the neural representations learned from raw waveform include more information for generating task-driven acoustic features. The best experimental result based on SincNet achieves an accuracy of 87%. This suggests that the approach of learning neural representations from raw waveform is capable of providing better performance than most models using domain-knowledge based acoustic feature sets/representations such as eGeMAPS, ComParE and BoAW representations in smoker identification tasks.

6. CONCLUSION

In this paper, we propose a dataset that can be used for the smoker identification study, and we investigate the efficiency of different acoustic features extracted/learned using three extraction/learning approaches for smoker identification. We find that all proposed acoustic features perform better than traditional conventional acoustic features (i.e., MFCC). To the best of our knowledge, this is the first study that comprehensively explores acoustic features for smoker identification from speech.

In the future, we will explore the effect of combining different acoustic feature sets/representations and also investigate the performance of using different deep neural networks as the classifiers. We will extend this study to learn how the acoustic properties of smokers' speech alter during the smoking cessation process (e.g., before they have fully stopped smoking, when they have quit smoking for one week, quit smoking for one month, etc.)

* ComParE-BoAW is reported here

7. ACKNOWLEDGEMENTS

This work is supported by the Performance-Based Research Fund (PBRF) Seeding Project, the University of Auckland and the 2020 Catalyst: Strategic NZ-Singapore Data Science Research Programme Fund, MBIE, New Zealand.

8. REFERENCES

[1] D. Pinar, H. Cincik, E. Erkul, and A. Gungor, "Investigating the Effects of Smoking on Young Adult Male Voice by Using Multidimensional Methods," *J. Voice*, vol. 30, no. 6, pp. 721–725, 2016.

[2] H. K. Ubhi, S. Michie, D. Kotz, O. C. P. van Schayck, A. Selladurai, and R. West, "Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features," *Transl. Behav. Med.*, 2016.

[3] A. H. Poorjam, M. H. Bahari, and others, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *2014 4th ICCKE*, 2014.

[4] C. H. Murphy and P. C. Doyle, "The effects of cigarette smoking on voice-fundamental frequency," *Otolaryngol. Neck Surg.*, 1987.

[5] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Dep. Linguist. Univ. Stock.*, 1994.

[6] J. Gonzalez and A. Carpi, "Early effects of smoking on the voice: A multidimensional study," *Med. Sci. Monit.*, vol. 10, no. 12, 2004.

[7] I. Guimarães and E. Abberton, "Health and voice quality in smokers: An exploratory investigation," *Logop. Phoniatr. Vocology*, vol. 30, no. 3–4, pp. 185–191, 2005.

[8] L. Lee, J. C. Stemple, D. Geiger, and R. Goldwasser, "Effects of environmental tobacco smoke on objective measures of voice production," *Laryngoscope*, vol. 109, no. 9, pp. 1531–1534, 1999.

[9] O. Zealouk, H. Satori, M. Hamidi, N. Laaidi, and K. Satori, "Vocal parameters analysis of smoker using

Amazigh language," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 85–91, 2018.

[10] A. Braun, "The effect of cigarette smoking on vocal parameters," *ESCA Work. Autom. Speak. Recognition, Identification, Verif. ASRIV 1994*, pp. 161–164, 2019.

[11] Z. Ma, C. Bullen, J. T. W. Chu, R. Wang, Y. Wang, and S. Singh, "Towards the Objective Speech Assessment of Smoking Status based on Voice Features: A Review of the Literature," *J. Voice*, 2021.

[12] A. H. Poorjam, S. Hesaraki, S. Safavi, H. van Hamme, and M. H. Bahari, "Automatic smoker detection from telephone speech signals," in *International Conference on Speech and Computer*, 2017, pp. 200–210.

[13] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 272–281, 2019.

[14] W. Xue, C. Cucchiarini, R. van Hout, and H. Strik, "Acoustic correlates of speech intelligibility. The usability of the eGeMAPS feature set for atypical speech," 2019.

[15] Y. Pan *et al.*, "Acoustic Feature Extraction with Interpretable Deep Neural Network for Neurodegenerative Related Disorder Classification.," in *INTERSPEECH*, 2020, pp. 4806–4810.

[16] J. Han *et al.*, "An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety," *arXiv Prepr. arXiv2005.00096*, 2020.

[17] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.

[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*, 2013, pp. 835–838.

[19] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," 2017.

[20] G. Gosztolya and R. Busa-Fekete, "Ensemble Bag-of-Audio-Words Representation Improves Paralinguistic Classification Accuracy," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 477–488, 2020.

[21] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth annual conference of the international speech communication association*, 2015.

[22] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, "End-to-end speech recognition from the raw waveform," *arXiv Prepr. arXiv1806.07098*, 2018.

[23] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[24] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion Identification from Raw Speech Signals Using DNNs.," in *Interspeech*, 2018, pp. 3097–3101.

[25] L. Brandschain, C. Cieri, D. Graff, A. Neely, and K. Walker, "Speaker Recognition: Building the Mixer 4 and 5 Corpora.," in *LREC*, 2008.

[26] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. of LREC*, 2010.

[27] F. Povolny *et al.*, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 75–82.

[28] N. Cummins, M. Schmitt, S. Amiriparian, J. Krajewski, and B. Schuller, "'You sound ill, take the day off': Automatic recognition of speech affected by upper respiratory tract infection," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017.

[29] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1–4, pp. 43–52, 2010.

[30] L. Zhang, J. Han, and S. Deng, "Unsupervised Temporal Feature Learning Based on Sparse Coding Embedded BoAW for Acoustic Event Recognition.," in *INTERSPEECH*, 2018, pp. 3284–3288.

[31] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech," *Proc. Interspeech 2021*, pp. 3415–3419, 2021.

[32] Y. Pan, V. S. Nallanthighal, D. Blackburn, H. Christensen, and A. Härmä, "Multi-Task Estimation of Age and Cognitive Decline from Speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7258–7262.

[33] H. Zeng *et al.*, "EEG emotion classification using an improved SincNet-based deep learning model," *Brain Sci.*, vol. 9, no. 11, p. 326, 2019.

[34] N. Cummins *et al.*, "A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition," in *Interspeech 2020*, 2020, pp. 2182–2186.

[35] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv Prepr. arXiv1607.06450*, 2016.

[36] Y. Laurent, C., Pereyra, G., Brakel, P., Zhang, Y., & Bengio, "Batch normalized recurrent neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016, pp. 2657--2661.

[37] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, 2013, vol. 30, no. 1, p. 3.