INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Detecting and analysing spontaneous oral cancer speech in the wild

Bence Mark Halpern[123], *Rob van Son*[13], *Michiel van den Brekel*[13], *Odette Scharenborg*[2]

[1]University of Amsterdam, ACLC, Amsterdam, The Netherlands
[2]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
[3]Netherlands Cancer Institute, Amsterdam, The Netherlands

b.halpern@nki.nl, r.v.son@nki.nl, M.W.M.vandenBrekel@uva.nl, o.e.scharenborg@tudelft.nl

Abstract

Oral cancer speech is a disease which impacts more than half a million people worldwide every year. Analysis of oral cancer speech has so far focused on read speech. In this paper, we 1) present and 2) analyse a three-hour long spontaneous oral cancer speech dataset collected from YouTube. 3) We set baselines for an oral cancer speech detection task on this dataset. The analysis of these explainable machine learning baselines shows that sibilants and stop consonants are the most important indicators for spontaneous oral cancer speech detection.

Index Terms: pathological speech, corpus, oral cancer speech, explainable AI

1. Introduction

Oral cancer is a disease which impacts approximately 529,500 people worldwide every year [1]. Apart from improving survival rates (*mortality*), research attention has shifted to improving the quality of life after surgery [2]. Oral cancer survivors can suffer from several problems affecting their quality of life: difficulty swallowing [3, 4], decreased tongue mobility [5] and impaired speech intelligibility [3]. The latter is the focus of our paper.

Speech impairment occurs due to the oral cancer treatment in which parts of the tongue or the entire tongue is removed (*partial/total glossectomy*). This partial or full removal causes an inability to reach articulatory targets. Oral cancer speech is consequently primarily impaired at the articulatory level, while only patients who have also undergone radiation therapy also have problems with phonation [3].

Different studies show different characteristics of oral cancer speech impairment. Stop consonants (mainly /k/, /g/, /b/, /p/, /t/, /d/) [6, 7] and alveolar sibilants (i.e., /s/, /z/) [8] seem to be primarily affected. In certain cases, patients are able to learn articulatory compensation techniques to adjust for the lost tongue tissue. For example, /t/ and /d/ can be produced by an altered bilabial seal [3]. The effect of glossectomy on vowels and diphthongs is less clear [9, 10].

Analysis of oral cancer speech has so far focused on read speech. In the studies above, participants were asked to read a text passage. However, it has been shown that such structured tasks can fail to identify some characteristics of speech [11].

So far, no research has been carried out investigating whether oral cancer speech can be reliably differentiated from non-oral cancer speech automatically. The aim of the paper is three-fold: 1) we investigate whether spontaneous oral cancer speech can be differentiated from healthy speech, focusing on spontaneous speech for the first time as far as we know, and as such present the first baselines for oral cancer speech detection. 2) In order to do so, we collected a large dataset, which allows us to use machine learning techniques. Creating a large dataset of pathological speech is time-consuming due to slow patient recruitment. We, therefore, created a database of "found" oral cancer speech, which is freely available to the community.[1] 3) We provide a preliminary analysis of the differences in spontaneous oral cancer speech and healthy speech.

Pathological speech detection is a broad field. There are two main approaches employed in the field. The first one is to develop a new acoustic feature using some knowledge about the pathological speech and use that in a simple classification model. Effectively, this is solving the problem in a divide-and-conquer approach: detecting known characteristics of a pathology and then feeding it into a classifier. A typical example of this is looking at unsuccessful phone realisations with an automatic speech recogniser (ASR) [12, 13]. The second approach is to generate some acoustic features from the audio using standard feature extractors (frontends) like openSMILE [14], Kaldi [15] or librosa [16]. This is a good approach if we are unsure what features would be the most appropriate. These features are then used to train a few chosen machine learning models (backends) such as artificial neural networks [17], Gaussian mixture models [18], support vector machines [19] and boosted regression trees [20]. These techniques rely on the models' learning capabilities to find any difference in the feature distributions of healthy and pathological speakers. We follow the second approach here, by using the Kaldi feature extractor along with ASR features.

In order to analyse the differences between oral cancer speech and healthy speech we only use backends which have some degree of explainability. Moreover, in addition to Kaldi features, we use phonetic posteriorgrams (PPG) as ASR-based features, which are easier to interpret than MFCC or PLP.

2. Dataset

We manually collected audio data containing English, spontaneous oral cancer speech from 3 male speakers and 8 female speakers from YouTube. The presence of oral cancer speech in the audio was determined by the content of the video and the authors' experience with such speakers. The audio was manually cut to exclude music, healthy speakers and artefacts, leaving only the oral cancer speech. The total duration of the oral cancer dataset is 2h and 59mins.

As our spontaneous healthy speech, we chose a subset of native American English speakers from the VoxCeleb dataset [21]. This dataset was chosen because it was also originally collected from YouTube. This allows exclusion of YouTube characteristics as a confounding factor in the detection task. The gender and number of speakers, as well as the amount of speech material for each speaker, was matched with that of the speech of the 11 oral cancer speakers to ensure that the ratio of the

[1]http://doi.org/10.5281/zenodo.3732322

http://dx.doi.org/10.21437/Interspeech.2020-1598

recordings is similar in the two datasets. There is no overlap in speakers between the training and test sets. In total, there are 10 speakers (8 female, 2 male) in the training set, and there are 12 (8 female, 4 male) speakers in the test set. The total duration of the training set is 4 h and 36 mins, for the test set 1 h 28 mins. The average duration per speaker is 27.6 min in the training set and 7.3 min in the test set.

The recordings in the oral cancer dataset were automatically cut into 5 s chunks to match the average duration of the utterances in the VoxCeleb dataset using `ffmpeg`. The audio was downsampled to 16 kHz and converted from stereo to mono. Loudness was normalised to 0.1 dB using the `sox` tool.

3. Method

We compared several frontend and backend combinations to find the best oral cancer vs. healthy speech detection system. Sections 3.1 and 3.2 describe the different frontends and backends, respectively, and the rationale why we chose them. The code of the analysis is also available online[2].

3.1. The preprocessing frontends

The following features have been extracted from the audio (abbreviations in bold). All features were calculated using the Kaldi frontend [15], unless mentioned otherwise. Silences were cut using Kaldi's voice activity detection (VAD) algorithm.

- **MFCC** - Mel-frequency cepstral coefficients are used as the baseline feature.
- **LTAS** - Long term average spectrum is used as a voice quality measurement in the early detection of pathological speech [22, 23] and to evaluate the effect of speech therapy or surgery on voice quality [24]. The LTAS features are extracted by calculating the mean and standard deviation of the frequency bins of Kaldi spectrograms and stacking them together.
- **PLP** - Perceptual linear predictive coefficients are known to be related to the geometry of the vocal tract based on the principles of source-filter theory [25]. During oral cancer the geometry of the vocal tract changes, so we expect that PLPs have useful information for detection. The PLPs are calculated based on [26].
- **Pitch** - To investigate whether there are also prosodic and phonation impairments in oral cancer speech, a combination of pitch and voicing likelihood feature is used [27].
- **PPG** - Phonetic posteriorgrams were calculated using an ASR trained on Librispeech [28], based on the implementation of [29]. PPGs are probability distributions over a set of phones, i.e., what is the probability that this phone is spoken at this frame of the utterance. The implementation that we used included 40 phones, including the phone for silence. However, silence phones were excluded in our approach.

3.2. The backends

Two different backends were used: a Gaussian Mixture Model (GMM) and a linear regression method (LASSO) [30]. Linear regression is generally the easiest to interpret, however when the dimensionality of the features are high, a feature selection step is usually recommended, that is why we used LASSO. The GMM is used widely in pathological speech detection [20, 18]. The GMM and LASSO models were implemented using the `sklearn` [31] library.

In addition to these two traditional models, we also trained a Dilated Residual Network (ResNet) [32]. Similar architectures have been successful in detecting spoofed speech [33, 34]. We expect ResNet's ability to recognise unnatural speech to be useful for detection of pathological speech.

3.2.1. Gaussian mixture model

We trained separate GMMs for oral cancer speech and healthy speech. The number of mixture components for each GMM was chosen so that it maximises performance on the test set from the list of $m = [4, 8, 10, 12, 16]$. This could result in overfitting to the test set, however, in practice we found that the test set performance is relatively insensitive to the choice of the mixture parameter. This is further discussed in Section 5. We report the number of mixture components used with the results in Table 1. At test time, we presented the healthy and the oral cancer speech utterances to both models. To determine whether the input speech frame contained healthy or oral cancer speech, we calculated the likelihoods for each speech frame and averaged over all frames to compute a single likelihood for the entire stretch of speech. The average likelihoods for both models were subsequently compared.

3.2.2. LASSO

LASSO is a variant of linear regression, which performs feature selection and regression simultaneously. It might be the case that for a given linear regression task, some features do not contain any relevant information to make predictions. In LASSO, coefficients of regression are encouraged to be close to zero if they do not provide useful information. Zeroing (pruning) some features means that the model requires only a subset of all predictors, making it parsimonious and easier to interpret. Pruning of the features is facilitated by setting the hyperparameter α: the larger this parameter is, the closer the coefficients are to zero. This hyperparameter is taken from the list $\alpha = [0.1, 0.01, 0.001, 0.0001]$ and tuned on the test set (see Section 5). The hyperparameters are reported with the results in Table 1.

3.2.3. Neural network classifier (Spectrogram-ResNet)

The ResNet architecture consists of four Dilated ResNet blocks. Each block has a different kernel size (width $\times$ height) and number of filters: (240×100) and 8; (120×200) and 16; (60×100) and 32; (30×50) and 64. This is followed by a fully connected layer with 100 hidden nodes and finally a softmax layer. The architecture is described in detail in in [33].

The input of the ResNet consisted of spectrograms. Spectrograms are highly informative, high dimensional features, which capture most properties of the raw speech signals. They are widely used with neural network backends [35, 34]. The input spectrograms were zero padded to the length of the longest utterance so that even the longest utterance could be processed by the network. The network was trained for 50 epochs in Keras [36], selecting the model with the best validation loss after training. We used the Adam optimiser with a learning rate of $\mu = 0.001$ [37]. To avoid overfitting on the test material, no hyperparameter optimisation was performed (see Section 5).

[2] `https://github.com/karkirowle/oral_cancer_analysis`

Table 1: *Equal error rates (EER) and accuracy of the classifiers with different feature and classifier combinations. Higher accuracy and lower EER is better. Best performances are emphasised in* ***bold****.*

GMM	**PLP**	**LTAS**	**PPG**	**Pitch**	**MFCC**	**Spectrogram-ResNet**
Train set accuracy	97.80%	94.71%	85.24%	52.04%	97.02%	98.58%
Train set EER	1.34%	5.3%	11.56%	39.07%	2.05%	1.00%
Test set accuracy	77.52%	65.59%	72.89%	43.57%	76.83%	**88.37**%
Test set EER	22.01%	31.05%	29.33%	45.65%	20.62%	**9.85**%
m	8	10	10	16	8	N/A
LASSO	**PLP**	**LTAS**	**PPG**	**Pitch**	**MFCC**	-
Train set accuracy	85.46%	98.55%	80.59%	70.22%	87.02%	-
Train set EER	9.25%	1.45%	12.25%	29.03%	8.01%	-
Test set accuracy	80.19%	87.37%	73.35%	58.86%	80.88%	-
Test set EER	20.62%	10.67%	25.84%	37.32%	19.23%	-
α	0.0001	0.01	0.001	0.001	0.01	-

4. Results and analysis

4.1. Results on training and test set

The detection accuracies for the training and test sets are measured using accuracy and equal error rate (EER), and can be seen in Table 1. Chance level accuracy for the test set is 57.82%. m refers to the number of Gaussian mixture components used during GMM training. α refers to the sparsity inducing hyperparameter of LASSO, a larger α indicates a sparser model.

The Spectrogram-ResNet-based detector achieved the best classification performance both in terms of accuracy and EER. This is closely followed by the LTAS-LASSO model. The superiority of the ResNet over the other methods is likely due to the ResNet seeing the whole utterance at once unlike the other methods. LASSO backends always outperformed the GMM-based backends on the test set, which is especially striking on the LTAS-based features. One possible explanation for the performance difference might be the collinearity of certain features as LASSO is known to handle collinearity better. We can see that for non-collinear features like MFCC or PLP the performance difference between LASSO and the GMMs is marginal. The worst performance was achieved by the Pitch features, which is close to chance level for both backends. This indicates that Pitch features are not appropriate for oral cancer speech detection, suggesting that oral cancer speech indeed is impaired primarily on the articulation level [3].

4.2. Analysis of the differences between oral cancer speech and healthy speech

To investigate the differences between oral cancer speech and healthy speech the two best performing architectures (Spectrogram-ResNET, LTAS-LASSO) and PPG-GMM were analysed through the information in the speech signal that was used by these models to distinguish oral cancer speech from healthy speech. While the PPG-GMM does not stand out in terms of accuracy, it lends itself to easy interpretation of the acoustic information used for the task.

4.2.1. Analysis of the Spectrogram-ResNet detector

To investigate what information the ResNet classifier uses to distinguish between oral cancer speech and healthy speech, we look at what parts of the spectrogram change the classification results the most.

To find these salient parts of the spectrogram, we calculate mean class activation maps, which indicate what frequencies are the most important for detection of both classes, for each sample in our test set as follows: Given a spectrogram image and a class label (oral cancer speech/healthy speech) as input, we pass the image through the ResNet to obtain the raw class scores before softmax. The gradients are set to zero for all classes except the desired class (i.e., oral cancer speech), which is set to 1. This signal is then backpropagated to the rectified convolutional feature map of interest [38]. We used the implementation from the `keras-vis` library.

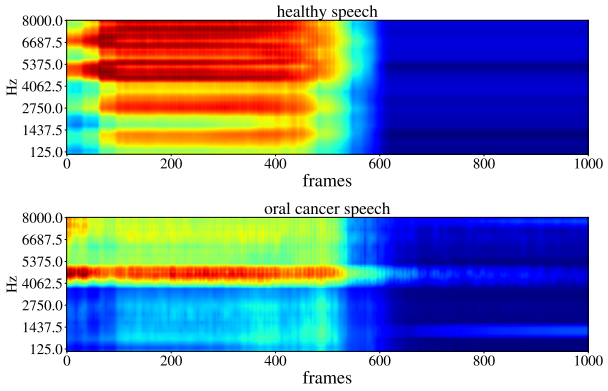

Figure 1: *Mean class activation maps for healthy speech (top panel) and oral cancer speech (bottom panel).*

Figure 1 shows the mean class activation maps for healthy speech and oral cancer speech. For healthy speech (top panel), the neural network spreads its focus (indicated by the coloured bands where red means higher intensity) among all the frequency bands. For oral cancer speech (bottom panel), the majority of the acoustic energy lies above the 4 kHz band. This indicates that sibilant frequencies, which are above 4 kHz [39], might be important for distinguishing between oral cancer and healthy speech. This is in agreement with previous studies who name sibilants as impaired sounds [8].

4.2.2. Analysis of the PPG-GMM detector

The trained GMM models can be viewed as models of a global oral cancer speaker and a global healthy speaker. The mean parameters of the GMMs can inform us what features are more typical for oral cancer speakers and which for healthy speakers by constructing a difference model. First, we calculate the difference of the mixture components in the two GMMs, obtaining a $d \in \mathbb{R}^m$ difference vector for each phone. Taking the mean of d, we are able to obtain a signed scalar $p \in \mathbb{R}$ for each phone class. If p is positive it means that there is a higher likelihood of occurrence of that phone in oral cancer speech compared to healthy speech. If p is negative it means that the likelihood of that phone is lower in oral cancer speakers – meaning that they have trouble pronouncing that phone.

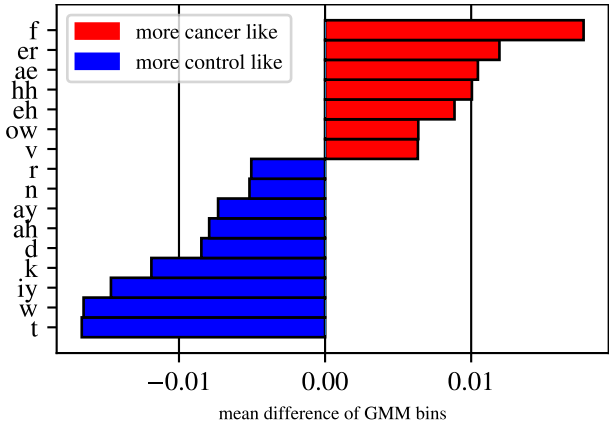

Figure 2: *Mean difference of GMM bins* (p) *of the PPG-GMM architecture.*

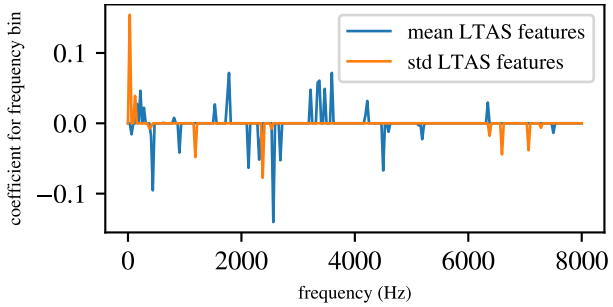

Figure 3: *Learned coefficients of the LTAS-LASSO.*

Figure 2 shows the results for the phones with absolute mean differences larger than 0.005. The blue bars indicate phones that are more often present in healthy speech and the red bars indicate phones that are more typical of cancer speakers. We can see that /t/, /w/, /iy/, /k/ and /d/ have lower likelihoods, indicating that some stop consonants are challenging for oral cancer speakers. This is in agreement with [6, 7].

4.2.3. Analysis of the LTAS-LASSO based detector

LASSO-based models can be analysed using the coefficients of regression. A positive coefficient indicates a feature contributing to the cancer class and vice versa. Figure 3 shows the learned coefficients. The blue line shows the mean coefficients, the orange line shows the standard deviation coefficients of the LTAS. Knowing that neighbouring frequencies are discouraged (because they provide similar (collinear) information), it is still surprising that some frequency bands have several adjacent positive/negative coefficients (clusters, shown as adjacent spikes). These clusters indicate that a greater level of frequency resolution is needed for that particular frequency band. We can see that for oral cancer speech this is the 3-4 kHz, indicating that sibilant frequencies need greater frequency resolution.

5. General discussion

The paper presents the first baseline models for the task of healthy vs. oral cancer speech classification. The Spectrogram-ResNet classifier achieved a high classification accuracy and outperformed all other models. A preliminary analysis of the three models indicated that sibilant frequencies, and stop consonant phones are the most decisive in the classification.

The current study used healthy speech from the VoxCeleb dataset, which only contains recordings from celebrities. Although potentially the detectors could use other features than those related to the acoustic characteristics of the speech for classification, this is not likely: inspection of the recognition results of the individual speakers in both datasets showed that in both datasets some speakers are well classified whereas others are not (range oral cancer speech: 49.7% – 100.0%; range healthy speech: 34.8% – 94.4% on the test set), although on average the oral cancer speakers were better classified than the healthy speakers: 89.6% vs. 66.2%.

Because of our relatively small-sized datasets, we kept hyperparameter tuning to a minimum to avoid overfitting on the test set. The more traditional methods (GMM and LASSO) only have a single hyperparameter, so chances of overfitting to the test set are small. Neural networks, on the other hand, usually have a myriad of hyperparameters, which makes overfitting more likely. To avoid this, we refrained from using any tuning mechanisms at all with the neural networks.

6. Conclusion

We presented a brand new dataset for analysis of spontaneous oral cancer speech, and showed that a detector based on ResNet taking spectrograms as input had a high performance in distinguishing between oral cancer speech and healthy speech, and generalised well to unseen data. Analysis of the speech signals through the classifiers shows that sibilants and stop consonants are important for oral cancer speech detection, while no evidence has been found on the importance of vowels and diphthongs.

7. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287. The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

8. References

[1] Kevin D Shield, Jacques Ferlay, Ahmedin Jemal, Rengaswamy Sankaranarayanan, Anil K Chaturvedi, Freddie Bray, and Isabelle Soerjomataram, "The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012," *CA: a cancer journal for clinicians*, vol. 67, no. 1, pp. 51–64, 2017.

[2] J B Epstein, S Emerton, D A Kolbinson, N D Le, N Phillips, P Stevenson-Moore, and D Osoba, "Quality of life and oral function following radiotherapy for head and neck cancer.," *Head Neck*, 1999.

[3] Elizabeth C Ward and Corina J van As-Brooks, *Head and neck cancer: treatment, rehabilitation, and outcomes. Chapter 5: Speech and Swallowing Following Oral, Oropharyngeal, and Nasopharyngeal Cancer*, Plural Publishing, 2014.

[4] Jeri A Logemann, Barbara Roa Pauloski, Alfred W Rademaker, and Laura A Colangelo, "Speech and swallowing rehabilitation for head and neck cancer patients," *Oncology*, vol. 11, no. 5, 1997.

[5] KDR Kappert, MJA van Alphen, LE Smeele, AJM Balm, and F van der Heijden, "Quantification of tongue mobility impairment using optical tracking in patients after receiving primary surgery or chemoradiation," *PloS one*, vol. 14, no. 8, 2019.

[6] Tim Bressmann, Hannah Jacobs, Janette Quintero, and Jonathan C Irish, "Speech outcomes for partial glossectomy surgery:

Measures of speech articulation and listener perception indicateurs de la parole pour une glossectomie partielle: Mesures de l'articulation de la parole et de la perception des auditeurs," *Head and Neck Cancer*, vol. 33, no. 4, pp. 204, 2009.

[7] Tim Bressmann, Robert Sader, Tara L Whitehill, and Nabil Samman, "Consonant intelligibility and tongue motility in patients with partial glossectomy," *Journal of Oral and Maxillofacial Surgery*, vol. 62, no. 3, pp. 298–303, 2004.

[8] Juha-Pertti Laaksonen, Jana Rieger, Jeffrey Harris, and Hadi Seikaly, "A longitudinal acoustic study of the effects of the radial forearm free flap reconstruction on sibilants produced by tongue cancer patients," *Clinical linguistics & phonetics*, vol. 25, no. 4, pp. 253–264, 2011.

[9] Juha-Pertti Laaksonen, Jana Rieger, Risto-Pekka Happonen, Jeffrey Harris, and Hadi Seikaly, "Speech after radial forearm free flap reconstruction of the tongue: a longitudinal acoustic study of vowel and diphthong sounds," *Clinical linguistics & phonetics*, vol. 24, no. 1, pp. 41–54, 2010.

[10] Marieke J De Bruijn, Louis Ten Bosch, Dirk J Kuik, Hugo Quené, Johannes A Langendijk, C René Leemans, and Irma M Verdonck-de Leeuw, "Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 180–187, 2009.

[11] Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp, "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *IEEE International Conference Mechatronics and Automation, 2005*. IEEE, 2005, vol. 3, pp. 1569–1574.

[12] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–6, 04 2008, Copyright - Copyright (c) 2008 S. Karger AG, Basel; Last updated - 2018-10-06.

[13] Andreas Maier, Elmar Nöth, Anton Batliner, Emeka Nkenke, and Maria Schuster, "Fully automatic assessment of speech of children with cleft lip and palate," *Informatica*, vol. 30, no. 4, 2006.

[14] Florian Eyben and Björn Schuller, "opensmile:) the munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[16] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.

[17] Philipp Klumpp, Julian Fritsch, and Elmar Nöth, "Ann-based alzheimer's disease classification from bag of words," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–4.

[18] Alireza A Dibazar, S Narayanan, and Theodore W Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*. IEEE, 2002, vol. 1, pp. 182–183.

[19] Tobias Bocklet, Andreas Maier, Josef G Bauer, Felix Burkhardt, and Elmar Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 1605–1608.

[20] Cassia Valentini-Botinhao, Sabine Degenkolb-Weyers, Andreas Maier, Elmar Nöth, Ulrich Eysholdt, and Tobias Bocklet, "Automatic detection of sigmatism in children," in *Third Workshop on Child, Computer and Interaction*, 2012.

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[22] Lindsey K. Smith and Alexander M. Goberman, "Long-time average spectrum in individuals with Parkinson disease," *NeuroRehabilitation*, 2014.

[23] Suely Master, Noemi De Biase, Vanessa Pedrosa, and Brasília Maria Chiari, "The long-term average spectrum in research and in the clinical practice of speech therapists," *Pro-fono : revista de atualizacao cientifica*, 2006.

[24] Kristine Tanner, Nelson Roy, Andrea Ash, and Eugene H. Buder, "Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy?," *Journal of Voice*, 2005.

[25] Gunnar Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 1, no. 1981, pp. 21–37, 1981.

[26] Hynek Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[27] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[29] Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Proc. Interspeech*, 2019, pp. 2843–2847.

[30] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[31] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[32] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser, "Dilated residual networks," *CoRR*, vol. abs/1705.09914, 2017.

[33] Rob van Son Anil Alexander Bence Mark Halpern, Finnian Kelly, "Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure," Odyssey 2020.

[34] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *Interspeech 2019*, ISCA, sep 2019, pp. 1013–1017, ISCA.

[35] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[36] François Chollet, "Keras: The Python Deep Learning library," *Keras.Io*, 2015.

[37] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[39] Peter Ladefoged and Sandra Ferrari DIsner, *Vowels and consonants*, Wiley & Blackwell, 2012.