Identifying Indicators of Vulnerability from Short Speech Segments using Acoustic and Textual Features

Xia Cui[1,2], Amila Gamage[2], Terry Hanley[1], Tingting Mu[1]

[1]The University of Manchester, United Kingdom
[2]VoiceIQ Ltd., United Kingdom

{xia.cui, terry.hanley, tingting.mu}@manchester.ac.uk, {xia, amila}@voiceiq.ai

Abstract

In order to protect vulnerable people in telemarketing, organisations have to investigate the speech recordings to identify them first. Typically, the investigation is manually conducted. As such, the procedure is costly and time-consuming. With an automatic vulnerability detection system, more vulnerable people can be identified and protected. A standard telephone conversation lasts around 5 minutes, the detection system is expected to be able to identify such a potential vulnerable speaker from speech segments. Due to the complexity of the vulnerability definition and the unavailable annotated vulnerability examples, this paper attempts to address the detection problem as three classification tasks: age classification, accent classification and patient/non-patient classification utilising publicly available datasets. In the proposed system, we trained three sub models using acoustic and textual features for each sub task. Each trained model was evaluated on multiple datasets and achieved competitive results compared to a strong baseline (i.e. in-dataset accuracy).

Index Terms: vulnerability detection, speech and text processing, age classification, accent classification, patient/non-patient classification, feature extraction, feature selection

1. Introduction

Protecting vulnerable people is a vital part of government regulation bodies and commercial companies in telemarketing [1, 2]. Vulnerability is a complex issue to detect as it is a multifaceted phenomenon that involves considering biological, psychological and social elements. According to [1, 2], everyone can be vulnerable – people with health conditions, older adults and children are arguably more likely to be vulnerable. Further, as the conversations conducted are conducted in English, people from non-English speaking countries may also be more vulnerable to being mis-sold products. When there is no priority information of the vulnerability criteria in the database, it is costly and time-consuming for an investigator to access a large number of recordings to identify the vulnerable people. Given this, there is an increasing demand to develop an automatic vulnerability detection system [2]. To the best of our knowledge, few studies [3, 4] have been conducted in the community on tackling the vulnerability in Speech Processing, addressing a fraction of vulnerability concerns. In order to adapt to a real-time system, this paper reports the development of a detection system that is able to work from a short speech segment (i.e. the average duration of an audio clip is less than 10 seconds). Without directly relying on the annotated vulnerability data to reduce product development cost, we propose a multi-task data-driven approach to detect the vulnerability through speech recordings by decomposing the task into a collection of sub-tasks, each of which can be solved by learning from publicly available data. Automatic

Speech Recognition (ASR) techniques have been greatly developed in the recent decade such as Deep Speech [5], we use both the speech transcriptions and acoustic waves to support the detection. More specifically, we investigated acoustic and textual feature selection that can be used for classifying speakers by age (i.e. child, adult or older adult), accent (i.e. native English speaker or non-native English speaker) and health status (i.e. patient with commonplace neurological difficulties or non-patient).

Our main contributions can be summarised as follows:

- We develop a vulnerability detection system for short speech segments with transcription by indicating the speaker's age group, accent group and health status.
- We study the feature extraction to detect the vulnerable people from speech segments. We found using a combination of acoustic and textual features works better than one modality (i.e. either speech or text) in most cases.
- Unlike prior research on investigating each feature, we investigate all possible combinations of feature groups.
- We evaluate three sub models on multiple benchmark datasets. Limited data resources are publicly available for evaluating the patient model, we collected and annotated a set of patient/non-patient speech segments accompanied with transcription from YouTube [1].

2. Related Work

Prior works demonstrated the potential of identifying vulnerable people such as patients with dementia [6], aphasia [7] and older adults [3, 4] using speech-based approaches. Feature extraction is an essential step to traditional approaches and deep learning approaches [8]. With the acquired popularity in ASR, extracting textual features from speech recordings along the acoustic features for speech classification become more reliable. In this section, we review some acoustic and textual features that have been frequently used in the prior works and can be applied to detect the vulnerability. Fundamental Frequency (F0) is a common measure for age and gender detection. Women have a higher F0 compared to men, and children have a higher F0 compared to adults [9, 10]. Spectral features such as Mel-frequency Cepstral Coefficients (MFCCs), Filter Bank Energies (FBEs) and Spectral Centroid Coefficients (SCCs) are frequently used in a number of applications [3, 6]. Voicing features such as the duration and number of unvoiced segments [6], and voiced utterances [9, 11] have shown the effectiveness in detecting language disorders. Jitter is a measure of frequency instability whereas shimmer is a measure of amplitude instability [9]. They are frequently used to detect the fluctuation and perturbation in speech signal respectively [12].

---

[1]https://www.youtube.com/

Harimonicity refers to Harmonic-to-Noise Ratio (HNR) and Noise-to-Harmonic Ratio (NHR) that measure the voice quality and are reported as a better measure for discriminating older adults and young people [13]. Mean of autocorrelation is another measure of voice quality estimating the pitch period of a given speech signal [6]. The Term Frequency-Inverse Document Frequency (TF-IDF) vectors are used to measure the repetitiveness by the cosine similarity between documents [14]. Part-Of-Speech (POS) features are represented by the frequency of various POS tags, such as interjections (i.e. filler words) were reported frequently in the use of detecting behaviour patterns and personality recognition [15]. Type Token Ratio (TTR) measures the weight of unique words in a document and shows the vocabulary richness (i.e. lexical diversity) of a document. A more advanced measure is moving-average type-token ratio (MATTR) [16], which computes the ratio by moving a fixed-size window within the document. Vulnerable people such as patients with memory problems and second language speakers are expected to have a lower TTR [11]. Psycholinguistic features were used for speech transcripts summarisation [17]. Older adults and people with certain health condition usually have memory problems. Several emotional categories from psycholinguistic features (e.g. depression, anxiety and stress) are often considered as causes for memory problems. In addition, the topical categories from psycholinguistic features can provide some insight to evidence on the speaker's life events.

3. Methods

We developed an automatic vulnerability detection system using a data-driven approach based on feature extraction and classification techniques. Below, we introduce the datasets and pre-processing (Section 3.1), feature extraction (Section 3.2) and classifier training details (Section 3.3).

3.1. Data

All sub-models were created and evaluated using features extracted from three English speaking TalkBank datasets (AphasiaBank [18], DementiaBank [19] and RHDBank [20]) and three large ASR datasets (Common Voice [2], VoxForge [3] and VCTK [4]). For ASR datasets such as Common Voice, we use the official validated subset. TalkBank datasets contain videos conducted and recorded by investigators and students, which are interviews with patients or people from health control group. The original video files were firstly converted into audio files via MoviePy [5]. Then, the audio files were trimmed into short clips by the timestamp and speaker label. In our scenario, models are created without any hand-crafted information, or probably based on the transcription from ASR. Therefore, the transcripts were downsampled. All hand-crafted information within the transcripts (e.g. timestamps for sub-sentences, POS tags, and manually-corrected words) were removed. Due to the recording devices, we found some audio files in the TalkBank are noisy, this was also reported in Al-hameed et al. [6]. Therefore, we used spectral gating [21] to reduce the stationary noise from the audio clips. Furthermore, we extracted available speaker information such as age, accent and gender for annotating the datasets. Depending on the model, we selected 1000 instances from each class to form a validation set for each model.

3.2. Feature Extraction

The feature extractor plays an important role in the system. Two sets of features we extract from recording and transcription are shown in Table 2. We implemented an acoustic feature extractor using parselmouth [6] and librosa [7]. Following Al-hameed et al. [6] and Teixeira et al. [22], we extracted acoustic features including 2 F0 variants (mean and covariance), first 42 MFCCs and their skewness, kurtosis, mean with kurtosis, mean with kurtosis and skewness of the mean, 26 FBEs, 26 SCCs, 5 pitch variants (mean, median, standard deviation, minimum and maximum), 4 pulses variants (number of pulses, number of periods, mean of and standard deviation of the periods), 3 voicing (fraction of locally unvoiced frames, number and degree of voice breaks), 5 jitter variants (local, local-absolute, the relative average perturbation, five-point perturbation quotient and the average absolute difference), 6 shimmer variants (local, local-dB, three point amplitude perturbation, five-point amplitude perturbation quotient ,eleven-point amplitude perturbation quotient and the average absolute difference) and 3 harmonicity variants (mean of the autocorrelation, NHR and NHR). We implemented a textual feature extractor using scikit-learn [8]. We extracted textual features including 3000 dimensional TF-IDF features, POS features, TTR and MATTR, psycholinguistic features and sentiment. We used the Universal POS tags [23] to form POS features, other POS tag marks such as Penn Treebank POS tags [9] can also be used. We use the pre-trained Convolutional Neural Networks (CNN) based sentiment analyser from stanza [24] to produce the sentiment feature. We use Empath [25] to extract a vector of 200 lexical categories to form the topic and emotion features.

3.3. Training

We address the vulnerability detection problem as three classification tasks to find the related indicators from speech recordings and corresponding transcriptions. A collection of three separate classification models were created: an age model to classify the speaker's age group (below 20 as child, between 20 and 60 as adult, and over 60 as older adult), a non-native model to classify the speaker's accent group (native and non-native English speaker) and a patient model to classify the speaker's health status (patient with aphasia, dementia or RHD and non-patient).

The number of instances used for training each sub model is summarised as (a) age model: child (14,472), adult (18,162) and older adult (9,943); (b) non-native model: native (31,500) and non-native English speaker (31,500); (c) patient model: patient (7,000) and non-patient (7,000). Due to incomplete speaker information available in the six datasets, we use different subsets for training different model. We employ a simple data fusion technique to combine multiple data sources in training. We learn a weight w_i for each training dataset, where w_i maximises the prediction accuracy on a validation set. The weights are learned using Bayesian Optimisation [26].

The age model is trained on a combination of six datasets: Common Voice, VCTK, VoxForge, AphasiaBank, DementiaBank and RHDBank. Their audio clips and corresponding transcripts are categorised into three age groups. We found the datasets are strongly imbalanced, we adjusted the class weight for training. In addition to the original age model, we train

Table 1: *Average classification accuracy (acc) with standard deviation (std), train and test time (in seconds) over 5-fold cross-validation using various learning algorithms for the age model using a vaildation dataset from Common Voice.*

Classifier	acc±std	Train Time	Test Time
Nearest Neighbors	0.6190±0.0227	1.2412	1.1662
Decision Tree	0.6390±0.0332	2.5636	0.1759
Random Forest	0.7565±0.0131	1.5734	0.2663
MLP	0.6060±0.0883	17.5251	0.1973
AdaBoost	0.7070±0.0058	15.5314	0.3440
Naïve Bayes	0.6110±0.0312	0.8323	0.2026
QDA	0.5090±0.0394	8.5438	0.9927
Logistic Regression	0.6885±0.0333	1.6408	0.1595
Linear SVM with SGD	0.5565±0.0446	1.2600	0.1667

a variant with separating training data into gender-age groups (e.g. female child, elderly male) and then map the gender-age groups back to their age groups. The non-native model is trained on a combination of three datasets: Common Voice, VCTK and VoxForge, of which the audio clips and transcripts are categorised into the two groups of native and non-native English speakers. We define native speakers by a list of native English speaking countries [10]. Unlike prior works that studied individual long-term illness and tried to differentiate patients from people in the health control group, we explored the possibility to discriminate the patients from non-patients in a more general sense. We train a patient model that aims at discriminating patients and non-patients by combining the three TalkBank datasets. In this scenario, we mix the data that are labelled as patients from the three datasets and use them as positive instances for training. Negative training instances are randomly selected from Common Voice.

4. Experiments

We conducted experiments on classification (Section 4.1), feature selection (Section 4.2) and evaluated each sub model on multiple datasets (Section 4.3).

4.1. Classification

To choose a proper classifier, we experimented several learning algorithms: Nearest Neighbors, Decision Tree, Random Forest, Multi-layer Perceptron (MLP), AdaBoost, Naïve Bayes, Quadratic Discriminant Analysis (QDA), Logistic Regression and Linear Support Vector Machine (SVM) with Stochastic Gradient Descent (SGD). We randomly selected 1000 older adult and 1000 non-older adult examples from Common Voice to train a binary classifier. We used the classifier implementation from scikit-learn [11]. Table 1 shows the average results using different classifiers over 5-fold cross validation. Random Forest classifier has shown improvements in a couple of speech classification tasks such as speech/non-speech discrimination [27] and speech emotion recognition [28]. We observed Random Forest classifier also shows a promising performance in speech age group classification and reaches competitive time efficiency in both training and testing.

4.2. Feature Selection

We conducted a comprehensive study into feature extraction and selection. More specifically, we run 5-fold cross-validation on each model over all possible combinations (32,767 combinations in total for 15 feature groups). Table 3 shows the top

---

Table 2: *Features with their dimensionality (dim). (·) denotes the shorthand name for each feature.*

	Features	dim
Acoustic Features	Mel-frequency Cepstral Coefficients (mfcc)	506
	Filter Bank Energy (fbank_energy)	26
	Spectral Centroid Coefficients (spectral_centroid)	26
	Fundamental Frequency (f0)	2
	Pitch (pitch)	5
	Pulses (pulses)	4
	Vocing (vocing)	3
	Jitter (jitter)	5
	Shimmer (shimmer)	6
	Harmonicity (harmonicity)	3
Text Features	TF-IDF (tf_idf)	3000
	Part-of-Speech Tags (pos_counts)	17
	Type Token Ratio (ttr)	2
	Topic and Emotion (empath)	194
	Sentiment (sentiment)	1

5 feature combination candidates for the age model and the top combination candidates are sorted by classification accuracy descendingly. In contrast to the prior works using either text [29] or audio features [4, 30] for estimating the age, we observe that most of the top candidates are combinations of both text and audio features (4 out of 5). Table 4 shows the ablation study of feature selection, the classification accuracy falls around 0.02 when we remove the text features such as TTR and sentiment. SCCs improve the performance significantly (i.e. around 0.07). Furthermore, we rank each feature by its occurrence in the top 10 combination candidates. Table 5 shows the frequently-occurred candidate features in top 10 combinations, the first row is the most frequently-occurred candidate feature and we add others to the following rows by their occurrence in the top 10 combinations. The results indicate audio features take a major role in the feature extraction for the age model. Using FBEs alone gains a good performance on the classification accuracy (0.574). MFCCs are frequently used as a promising feature for the audio classification, however, we find using MFCCs alone achieves around 0.59 in accuracy, which is more computational costly (i.e. the dimensionality of MFCCs is 506) and does not perform as well as a combination of the other audio features with lower dimensionality. In our preliminary experiments, we found TF-IDF feature had a strong impact on the performance. TF-IDF usually fails if a test sentence contains many out-of-vocabulary words. To expand the feature space to overcome this issue, one of the possible solutions is to use word embeddings. Table 6, we compare the TF-IDF feature (tf_idf), topic and emotion feature (empath) with some popular word embeddings such as Fast-Text (crawl and news) [31], Extended Dependency Skipgram (extvec) [32], GloVe (glove) [33], Skip-gram (twitter) [34] and Turian (turian) [35]. We use the implementation from flair [36] and each sentence is represented by a fix-length 100 dimensional embedding. The age model is a three-class classifier, both TF-IDF and word embeddings do not improve the classification accuracy significantly (i.e., close to 0.3333). Due to the page limit, we presented the results on one of the sub models, similar trend is also observed in the other models.

4.3. Model Evaluation

Table 9 shows the average classification accuracy of three sub models with an additional age model variant evaluated on multiple datasets. We first evaluate the trained age model on all six datasets. Table 7 shows the classification accuracy on each dataset's test set. In-dataset accuracy denotes the classification accuracy using the given dataset, and it is often considered as a

Table 3: *Top 5 feature combination candidates for the age model with the acc and std on 5-fold cross validation, sorted by the accuracy descendingly.*

Features	acc±std	dim
pulses + harmonicity + fbank_energy + spectral_centroid + f0 + sentiment + ttr	0.662± 0.023	64
pitch + voicing + jitter + harmonicity + fbank_energy + spectral_centroid + f0 + ttr	0.660± 0.019	72
pitch + fbank_energy + spectral_centroid + sentiment	0.660± 0.014	58
pitch + pulses + harmonicity + fbank_energy + spectral_centroid + f0	0.660±0.017	66
pitch + harmonicity + fbank_energy + spectral_centroid + f0	0.659±0.018	62

Table 4: *Ablation study of the feature combination (age model).*

Features	acc±std	dim
pulses + harmonicity + fbank_energy + spectral_centroid + f0 + sentiment + ttr	0.662±0.023	64
(-ttr)	0.648±0.021	62
(-sentiment)	0.642±0.016	63
(-f0)	0.650±0.020	62
(-spectral_centroid)	0.597±0.019	38
(-fbank_energy)	0.624±0.017	38
(-harmonicity)	0.647±0.023	61
(-pulses)	0.646±0.027	60

Table 5: *Frequently-occurred candidate features (age model).*

Features	acc±std	dim
fbank_energy	0.574±0.013	26
fbank_energy + spectral_centroid	0.623±0.023	52
fbank_energy + spectral_centroid + harmonicity	0.636±0.010	55
fbank_energy + spectral_centroid + harmonicity + f0	0.648±0.024	57
fbank_energy + spectral_centroid + harmonicity + f0 + pitch	0.659±0.018	62
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses	0.660±0.017	66
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment	0.648±0.014	67
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment + ttr	0.653±0.012	69
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment + ttr + voicing	0.637±0.024	72
fbank_energy + spectral_centroid + harmonicity + f0 + pitch + pulses + sentiment + ttr + voicing + jitter	0.655±0.014	77

Table 6: *Evaluation on embedding features (age model).*

Embedding	crawl	extvec	glove	news
acc+std	0.3647±0.0266	0.3690±0.0191	0.3643±0.0087	0.3473±0.0302
Embedding	turian	twitter	tf_idf	empath
acc+std	0.3643±0.0284	0.3507±0.0203	0.3530±0.0128	0.3393±0.0182

Table 7: *Classification accuracy tested on various datasets for the age model. #test denotes the number of test examples.*

Dataset	#test	In-dataset	Data Fusion	Data Fusion + Gender Separation
Common Voice	9000	0.7249	0.7414	0.7460
VoxForge	3579	0.8178	0.8128	0.8268
VCTK	4000	0.9423	0.9398	0.9480
AphasiaBank	1099	0.6016	0.6497	0.6261
DementiaBank	73	0.8493	0.7808	0.7945
RHDBank	772	0.9391	0.9443	0.9313

Table 8: *Classification accuracy tested on various datasets for the non-native model.*

Dataset	#test	In-dataset	Data Fusion
Common Voice	18000	0.8066	0.8042
VoxForge	6000	0.860	0.8584
VCTK	3000	0.967	0.9553

Table 9: *Average classification accuracy evaluated on multiple datasets for all trained models with top feature combination and dimensionality.*

Model	acc	Top Feature Combination	dim
Age (+Gender Separation)	0.8121	jitter + shimmer + fbank_energy + spectral_centroid + f0 + sentiment + ttr	64
Age	0.8115	pulses + harmonicity + fbank_energy + spectral_centroid + f0 + sentiment + ttr	68
Non-Native	0.8726	pitch + voicing + jitter + shimmer + fbank_energy + spectral_centroid	71
Patient	0.6840	shimmer + harmonicity + mfcc + fbank_energy + spectral_centroid + f0 + pos_counts + ttr	588

curacy across these datasets. With the help of gender separation, 4 out of 6 datasets perform slightly better than the original age model. Table 8 shows the classification accuracy for the non-native model tested on ASR datasets. Common Voice contains a large number of non-native speakers that are Indian, whereas VoxFroge contains a large number of non-native speakers are European. VCTK was claimed as a native English speaker dataset in the original publication. However, we found it contains one speaker from India. Considering the diversity of the three datasets, the proposed data fusion model is relatively robust that the test accuracy is slightly lower than the in-dataset accuracy. For evaluating the patient model, we retrieved 59 videos from YouTube using some patient related keywords: *aphasia+example*, *dementia+example*, *mental+ill+patient* and *patient+voices+nhs*. "+" denotes an AND relation in a search query. We also retrieved the associated transcription. We converted these videos into audio and trim them into short clips by the timestamp in the transcription. However, no speaker information is available for this dataset. We randomly select 510 audio clips from this dataset and manually annotate them as patient (124) or non-patient (386) voice clips. Due to the transcription quality, this model was evaluated under a challenging noisy condition and the results can be treated as a baseline for future development. The patient model achieved a classification accuracy of 0.684 (Table 9). We observed a relatively low false positive rate (0.1891) but a low true positive rate (0.2903).

5. Conclusion and Future Work

We studied the features extracted from short speech segments and their transcription. We address the detection problem by dividing it into three separate classification tasks: age classification, accent classification and patient/non-patient classification. We trained an age model, a non-native model and a patient model respectively. We evaluated the age and non-native models on multiple benchmark datasets. The patient model was evaluated on a manually-annotated dataset collected from YouTube. We presented a data-driven approach to address the vulnerability detection problem. The models were trained using supervised learning algorithms on extracted features. We plan to extend this work using semi-supervised learning and pre-trained deep learning models for speech to reduce the number of labelling data required for training. In the future, we will adapt the vulnerability detection system to the practice.

strong baseline for evaluating the model generalisation. Common Voice and VoxForge contain data from all three age groups. We observe a slight improvement by data fusion and gender separation compared to the in-dataset accuracy. VCTK does not contain any data from older adult class and the age range is narrow (speakers are 18 to 30 years old). In this case, a binary in-dataset classifier is trained. AphasiaBank, Dementia-Bank and RHDBank have a similar situation that there is no or few data from the child class and the age range is close to the pre-defined boundary. We observe the proposed model still performs competitively under this challenging condition. In general, by using data fusion to introduce additional data sources, a few improvements can be observed in the classification ac-

6. References

[1] "Consumer vulnerability," Financial Conduct Authority (FCA), Tech. Rep. 8, Feb. 2015.

[2] "Consumer vulnerability: challenges and potential solutions," Competitions & Markets Authority (CMA), Tech. Rep., Feb. 2019.

[3] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[4] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[6] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for alzheimer's disease," in *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 32–36.

[7] A. Shivkumar, J. Weston, R. Lenain, and E. Fristed, "Blabla: Linguistic feature extraction for clinical analysis in multiple languages," *arXiv preprint arXiv:2005.10219*, 2020.

[8] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[9] C. L. Lortie, M. Thibeault, M. J. Guitton, and P. Tremblay, "Effects of age on the amplitude, frequency and perceived quality of voice," *Age*, vol. 37, no. 6, p. 117, 2015.

[10] A. Heidari, A. Moossavi, F. Yadegari, E. Bakhshi, and M. Ahadi, "Effects of age on speech-in-noise identification: subjective ratings of hearing difficulties and encoding of fundamental frequency in older adults," *Journal of audiology & otology*, vol. 22, no. 3, p. 134, 2018.

[11] M. Yancheva, K. C. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 134–139.

[12] M. L. B. Pulido, J. B. A. Hernández, M. Ángel Ferrer Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: A review," *Expert Systems with Applications*, vol. 150, p. 113213, 2020.

[13] C. T. Ferrand, "Harmonics-to-noise ratio: an index of vocal aging," *Journal of voice*, vol. 16, no. 4, pp. 480–487, 2002.

[14] V. Masrani, G. Murray, T. S. Field, and G. Carenini, "Domain adaptation for detecting mild cognitive impairment," in *Canadian Conference on Artificial Intelligence*. Springer, 2017, pp. 248–259.

[15] F. Alam and G. Riccardi, "Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 955–959.

[16] M. Covington and J. D. McFall, "The moving-average type-token ratio," *Linguistics Society of America, Chicago, IL*, 2008.

[17] S. K. Barnwal and U. S. Tiwary, "Using psycholinguistic features for the classification of comprehenders from summary speech transcripts," in *Intelligent Human Computer Interaction*, P. Horain, C. Achard, and M. Mallem, Eds. Cham: Springer International Publishing, 2017, pp. 122–136.

[18] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[19] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[20] B. MacWhinney, "Understanding spoken language through talkbank," *Behavior research methods*, vol. 51, no. 4, pp. 1919–1927, 2019.

[21] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[22] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[23] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," in *Proceedings of the International Conference on Language Resources and Evaluation*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2089–2096.

[24] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of ACL: System Demonstrations*, 2020.

[25] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4647–4657.

[26] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *NIPS*. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 2951–2959.

[27] S. V. Thambi, K. Sreekumar, C. S. Kumar, and P. R. Raj, "Random forest algorithm for improving the performance of speech/non-speech detection," in *2014 First International Conference on Computational Systems and Communications (ICCSC)*. IEEE, 2014, pp. 28–32.

[28] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[29] R. G. Guimaraes, R. L. Rosa, D. De Gaetano, D. Z. Rodriguez, and G. Bressan, "Age groups classification in social network using deep learning," *IEEE Access*, vol. 5, pp. 10 805–10 816, 2017.

[30] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors." in *INTERSPEECH*, 2016, pp. 1402–1406.

[31] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[32] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in *NAACL-HLT*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1490–1500.

[33] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[35] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 384–394.

[36] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.