Voice Technology to Identify Fatigue from Japanese Speech

Raymond Brueckner[1], *Misa Takegami*[2], *Namhee Kwon*[1], *Nate Blaylock*[1], *Vinod Subramanian*[1], *Eri Kiyoshige*[2], *Soshiro Ogata*[2], *Yuriko Nakaoku*[2], *Henry O'Connell*[1], *Kunihiro Nishimura*[2]

[1]Canary Speech, USA, [2]National Cerebral and Cardiovascular Center, Japan

{ray, namhee, nate, vinod, henry}@canaryspeech.com
{takegami, kiyoshige.eri, s_ogata, yurikon, knishimu}@ncvc.go.jp

Abstract

Toward an automatic health monitoring tool, we investigate voice analysis technology to extract related features from speech and to build a machine learning model for identifying fatigue in Japanese. We collect voice data and their fatigue labels through phone calls and then experiment with diverse machine learning methods using various acoustic and prosodic features. The models are trained on spontaneous Japanese speech from participants who are older than 70 years. Each model and feature shows different performance and the logistic regression model using x-vectors trained on English outperforms other models with sensitivity at 0.87 and specificity at 0.65.

1. Introduction

The ability to monitor health of elder adults in a home environment is an important key to supporting aging in place, as well as handling the increasing needs of elder care, especially in countries where there is an aging population. Vocal biomarkers represent an automatic, non-invasive and objective way to monitor and screen for diseases and other human conditions at home. In this context screening for fatigue is of high clinical relevance.

In this study, we focus on vocal biomarkers for detecting fatigue as measured by the Short Form 36 (SF-36) health questionnaire vitality score [1]. The vitality section of the SF-36 questionnaire captures the energy and fatigue aspects of health [1]. Fatigue can refer to a variety of conditions such as sleep fatigue [2], work fatigue [3, 4, 5] and COVID-19 fatigue [6, 7]. According to Ware et al. [1], the vitality part of the questionnaire was added to capture "differences in subjective well-being". This 'vitality' measure of fatigue is important, especially for older adults, as it has been linked to an increased risk for various issues including falls [8], heart disease [9], and even mortality [10].

Although SF-36 is a reliable approach to measuring a health condition [11], the cost of administering SF-36 is high [12] and repetitive surveys are potentially inefficient to monitor health conditions. We explore non-invasive voice processing technology to consistently monitor individuals' health, especially of older adults. By building a speech-processing tool, we aim at identifying issues or changes in health and providing proper actions in time.

In this work, we train different machine learning models to analyze features derived from voice recordings of senior citizens in Japan and predict whether they show signs of fatigue indicated on the SF-36 vitality measurement.

To our knowledge, no direct research using audio to detect fatigue based on the SF-36 questionnaire has been done so far. In the area of sleep fatigue detection, the authors of [2] collected data from 15 participants and a Support Vector Machine (SVM) was trained on common audio features such as F_0, Mel-Frequency Cepstral Coefficients (MFCCs), and loudness. Similarly, for work fatigue detection Krajewski et al. [3] collected data from 12 participants and common audio features were used to train an ensemble of machine learning models such as a linear SVM, logistic regression, decision trees, and multi-layer perceptrons. COVID-19 fatigue detection was investigated by Han et al. [6] on a dataset consisting of 52 participants, where they used common audio features to train an SVM.

Across existing fatigue research, the usual approach is to use common audio features to train a shallow machine learning classifier. Further, the detection of fatigue is typically treated as a binary classification problem. The shortcoming of pre-existing work is that adopted datasets are usually quite small and clear details about how the data are split between training and testing sets are not precisely specified.

Berisha et al. [13] showed that models built on larger amounts of data tend to report lower accuracy than the models using fewer data in the clinical speech machine learning literature, which can be interpreted as overfitting on smaller data sets. For this reason, we collect a comparably large data set for the prediction of fatigue consisting of more than 1500 samples to build reliable models and we compare a variety of machine learning approaches to this task. Furthermore, this study is conducted on Japanese speech and our model uses language-agnostic acoustic features. The final prediction unweighted accuracy on fatigue is 0.76 and its sensitivity and specificity are 0.87 and 0.65, respectively.

2. Data Collection

2.1. Phone Survey

We collected speech audio samples and self-answered labels through phone interviews in Japan with senior citizens beyond the age of 70 years. Trained representatives called the randomly selected elderly adults to ask them a catalog of questions if they agreed on the data collection. Previous research on anxiety and depression had shown that self- and telephone-administered assessments can be used interchangeably [14] as well as in-person interviews [15]. The representatives were given the interview scripts including the list of self-evaluation questionnaires and the example questions to elicit free speech. The free speech questions include topics on shopping, housework, job, etc.

2.2. Fatigue Label

A Japanese version of the SF-36 (36-item Short Form Survey) questionnaire [16] was asked during the phone call, and the results of the vitality (energy and fatigue) subsection were used

as the fatigue label. The vitality score ranges between 0 and 100, where the lower score means more fatigued. The representatives wrote down the interviewee's answers to the itemized questions about vitality. Later, we generated a binary label *Fatigue* by cutting off the vitality score at the 25 percentile (at the score of 50).

2.3. Data Pre-Processing

We collected the audio files in the uncompressed WAV format in order to reduce any potential negative effect of compression on the final prediction of our models. Since they were collected through conversations over the phone[1], the sampling rate was set to 8 kHz. We obtained de-identified participant channel audio and performed segmentation by considering pause duration and the remaining segment duration. We filtered out all audio segments representing answers to vitality questionnaires and instead used only free-speech portions. The filtering was done by several heuristic rules on the keywords and segment duration. In this study we focus on the analysis of free speech instead of answers to questionnaire questions since our product goal is to build a voice analysis tool for early detection of fatigue from general speech.

The selected audio segments were concatenated resulting in one audio sample per phone call. We measured the duration of voice activity by adding the duration of each spoken word as determined by ASR (automatic speech recognition), excluding pauses between words. Although we use full audio samples including pauses for our analysis, we filtered out short audio samples based on the voice activity duration. The statistics over each sample's voice activity duration are presented in Table 1. Since very short speech samples - the minimum voice activity duration d_{min} is just about 3 seconds - do not contain sufficient information to robustly detect fatigue from speech, samples whose voice activity duration was shorter than 20 seconds were removed from the data set. Out of the 1535 phone calls (samples) that were collected, 236 short audio files were excluded, so the final number of samples we used for our study was 1299.

Table 1: *Statistics of Sample-Wise Voice Activity Duration*

Statistical Value	Voice Activity Duration [sec]
Mean	69.3
StdDev	49.2
Minimum	3.1
Maximum	429.8
Voice Activity Duration	Number of Samples
$\geq$ 40 seconds	1035
$\geq$ 30 seconds	1196
$\geq$ 20 seconds	1299
Total collected	1535

3. Acoustic Voice Features

From the pre-processed audio samples, we generated various feature sets, both hand-engineered and data-driven, which serve the subsequent classification and prediction step as an intermediate input representation. The rationale for this commonly adopted step in a classification system is three-fold: first, it serves to extract relevant and eliminate irrelevant information. Second, it provides a compact representation of the underlying audio recording, constituting a dimensionality reduction mechanism. And finally, it transforms the variable-length audio signal into a fixed-sized intermediate representation required by most classification algorithms. Generally, acoustic features are different from lexical features in that they are language-agnostic to a large extent. Yet, especially data-driven features might still exhibit some language dependency depending on the data they were derived from.

In this study, we compare a number of different feature sets whose dimensionality (per speech sample) is depicted in Table 2.

Table 2: *Feature sets*

Features	Dimensionality
Acoustic	1944
Prosodic	33
i-vector (Japanese)	512
x-vector (Japanese)	512
x-vector (English)	512
openSMILE (eGeMAPSv02)	88
openSMILE (Speaker Trait)	5757
openSMILE (ComParE 2016)	6373

These feature sets are defined as follows:

- **Acoustic** features are calculated on a per-frame basis, where frames are defined as 25 ms sliding windows over the audio signal being computed every 10 ms. For each window, a 36-dimensional vector is calculated consisting of Mel-Frequency Cepstral Coefficients (MFCC) [17], Perceptual Linear Prediction (PLP) [18], and fundamental frequency (F_0) features. From the sequence of these vectors, its first-order and second-order time derivatives are derived and added to the base features. Based on a word-boundary Voice Activity Detector (VAD) of an ASR system, all non-speech feature segments are subsequently removed. Finally, 18 statistical values such as mean, percentile, slope, skewness, kurtosis, quartile, etc., are computed for each individual feature component across all (speech) frames.
- **Prosodic** features capture aspects of the modulations caused by the movement of the human articulatory organs during speech. They mainly consist of statistics of the fundamental frequency (F_0), energy, and rate of speech, derived via supra-frame spectral analysis. In particular, *normalized deciles* are computed by normalizing the deciles of F_0 and energy values of a given speech signal by its first decile, as given by

$$\gamma_i = \log(\frac{\phi_i}{\phi_1}), i \in \{2, 3, 4, \cdots, 9\} \quad (1)$$

  where ϕ_i denotes i-th decile. Besides, the minimum and maximum values of the F_0 and energy over the full audio signal are added to the feature set. In addition, as a measure related to speech rate the rhythm of the acoustic energy pattern is analyzed, resulting in features representing the duration and number of syllables, pauses, speech segments, phonation, articulation, and average speaking duration (ASD).
- **I-vectors** are data-driven low-dimensional speech representations based on factor analysis using Gaussian Mixture Mod-

[1]We ignore the effect of any compression in the phone transmission channel, since this is not under our control.

els (GMM) [19] and were originally invented for the purpose of robust speaker recognition [20]. They effectively constitute a fixed-sized representation of a variable-length audio signal and are trained on a data corpus in a supervised fashion. Since no pre-trained 8 kHz models are publicly available, we trained an i-vector model on the *Corpus of Spontaneous Japanese (CSJ)*[2] [21] and the *Callhome Japanese* corpus [22].

- **X-vectors** were proposed as an improvement over i-vectors and are built using a deep neural network (DNN) [23]. They, too, are trained in a supervised, data-driven fashion and return a fixed-sized representation of a variable-length input. In our study we used a publicly available English x-vector model[3] trained on 8 kHz audio data from several data sets including *Switchboard* [24], *Mixer 6* [25] and *NIST SREs* [26]. In addition, we trained our own x-vector model for Japanese on the *CSJ* and *Callhome Japanese* corpora, similar in spirit as for the i-vector model above.
- **openSMILE** features originate from the research area of voice-based Paralinguistics [27]. In our study, we investigate three different feature configurations: eGeMaps [28], ComParE 2016 [29], and Speaker Trait [30], each computed using the openSMILE 3.0 package [27]. These features are derived from frame-based, low-level audio descriptors, similar to the acoustic features described above, but undergo an additional application of a large number of functionals (e. g. moments, extrema, percentiles, peaks, and valleys, etc.), resulting in a single feature vector per audio signal [31].

4. Model Architectures

We built binary classification models predicting the fatigue label as defined by the vitality score in SF-36. Various model architectures were explored including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), as well as a fully connected deep neural network (DNN-FC).

The fully connected DNN model was built using Tensorflow [32] adopting hyperband search [33] to tune the hyperparameters such as layer depth, hidden layer size, and drop-out rate while monitoring the performance on the development set. We used the Adam optimizer with an *inverse time decay* learning rate and a batch size of 16 for 100 epochs with early stopping. Each layer applies a *Rectified Linear Unit (ReLU)* activation function, except for the last layer, which uses the *Sigmoid* function. The remaining (shallow) models were implemented using scikit-learn [34], namely an SVM with a *Radial Basis Function (RBF)* kernel and a Random Forest classifiers consisting of 100 trees in the forest. In order to reduce the data imbalance, all models apply class weights computed from the prior class distributions of the training data.

5. Experiments

5.1. Training

As described in Section 2.3, we used 1299 samples whose voice activity duration is longer than 20 seconds and split the data randomly into 3 disjunct groups, namely training, development and test, with each speaker belonging to only one of the data groups. The distribution over the number of samples for each group is shown in Table 3. We trained various models on the training set by tuning the hyperparameters using the development set and reported the final performance on the prediction result on the test set.

Table 3: *Data Split Distribution*

Data Set	Fatigue	Non-Fatigue
Train	306	829
Dev	21	63
Test	15	65

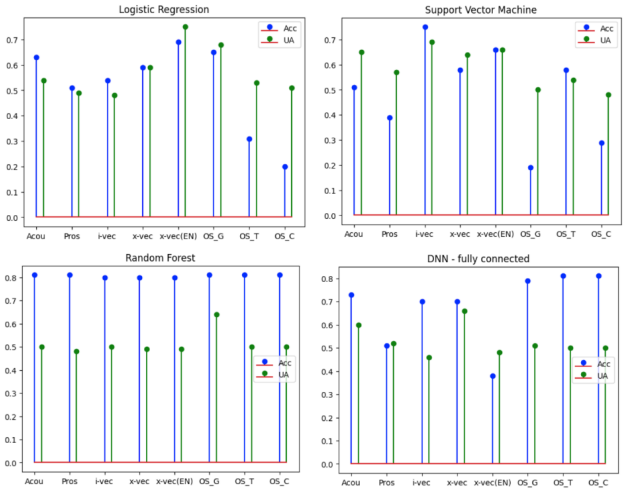
5.2. Evaluation Metrics

We measured the sensitivity and specificity along with the accuracy (Acc). The sensitivity is the correct prediction rate of the positive label (fatigue) and the specificity is the correct prediction rate of the negative label (non-fatigue). We also report the unweighted accuracy (UA), which is the average of sensitivity and specificity, i. e. the respective class recalls. One benefit of reporting the unweighted accuracy is that we can perform a fair comparison when the label distribution is imbalanced, because the model's chance UA is always 0.5 regardless of the label distribution. Similarly, we use the Area-Under-The-(ROC)-Curve (AUC) to measure the models' performance considering the trade-off between sensitivity and specificity based on a threshold on the prediction probability.

5.3. Results

We report the classification model performance for each model architecture and feature set. Figure 1 shows the accuracy and the unweighted accuracy per model.

Figure 1: *Model Performance per Feature*



The logistic regression model using the x-vector (x-vec) features showed the highest UA. Interestingly, its x-vector features were pre-trained on English corpora and it even outperformed the models using x-vectors pre-trained on Japanese corpora. The same behavior was observed in the support vector machine, but not in the DNN model. It could be related to the fact that the number of speakers in the English training corpora

[2] https://clrd.ninjal.ac.jp/csj/en/index.html
[3] https://kaldi-asr.org/models/3/0003_sre16_v2_1a.tar.gz

Table 4: *Best Model Performance for each Feature set*

Features	Model Architecture	Accuracy	UA	Sensitivity	Specificity	AUC
Acoustic	fully-connected DNN	0.73	0.60	0.40	0.80	0.62
Prosody	fully-connected DNN	0.51	0.52	0.53	0.51	0.53
i-vector (JP)	Support Vector Machine (SVM)	0.75	0.69	0.60	0.78	0.69
x-vector (JP)	fully-connected DNN	0.70	0.66	0.60	0.72	0.72
x-vector (EN)	Logistic Regression (LR)	0.69	0.76	0.87	0.65	0.76
openSMILE (eGeMaps)	Logistic Regression (LR)	0.65	0.68	0.73	0.63	0.66
openSMILE (Speaker Trait)	Support Vector Machine (SVM)	0.59	0.54	0.47	0.62	0.54
openSMILE (ComParE 2016)	Random Forest (RF)	0.82	0.63	0.33	0.93	0.68

is larger than the number of speakers in the Japanese training corpora, so the generated x-vectors are presumably more reliable. The support vector machine model using i-vectors performed as well as the x-vector logistic regression model and showed higher accuracy and lower UA. The DNN model's accuracy was decently high for acoustic and x-vector features. However, the random forest model failed to perform well for most feature sets. The openSMILE (OS) features performed differently depending on the feature configuration, and the eGeMAPs configuration (OS_G) performed higher than the other configured openSMILE feature sets.

The best-performing model architectures and their evaluation results for each feature set are shown in Table 4. The logistic regression model using an English x-vector achieved an accuracy of 0.69 and an unweighted accuracy of 0.76, where the sensitivity and specificity are 0.87 and 0.65 respectively. Since our goal is to identify people suffering from fatigue, we prefer the best sensitivity model, however, the support vector machine model using i-vector is also promising by showing higher accuracy (0.75) and higher specificity (0.78).

One of the challenges of this experiment is extracting features from phone conversations that often contain situations of distraction and many short answers. Although the representatives ask questions to elicit free speech, the answers vary for each individual in terms of audio length and answer quality. To obtain long enough voice data in our experiment, we concatenated several short segments as input data (described in Section 2.3), which could potentially add artifacts on the linked part of two segments. In the future, we will explore how to analyze segments with various lengths separately and together while minimizing the input artifacts.

Our experiment result shows that each model's performance strongly varies across the respectively adopted model architecture given a specific feature set. Identifying fatigue from speech is not a straightforward task because the acoustic patterns are sparse. Given the sparse features, the model performance is much affected by the training data set and model architecture. We conjecture that more data would help to learn all the patterns more robustly. Further, based on the fact that the x-vector model pre-trained on the English corpora performed well, we will investigate the more advanced domain adaptation methods where we can use the English resource efficiently.

Yet the experiments of this study show that we are able to achieve an encouraging result in predicting a subjective fatigue label from noisy speech data. We will explore how to analyze best the actual conversations and interactions in phone calls and how to apply voice analysis technology to self-assessment answers in a mobile application.

6. Conclusions

In this study, we investigated different machine learning model architectures to identify fatigue from spontaneous, free speech exploring various different acoustic, prosodic, and data-driven features, namely logistic regression, support vector machines, random forests, and a fully-connected DNN models. We found that not all models and feature combinations yield high-accuracy predictions, but by adopting a logistic regression model using x-vectors pre-trained on English corpora, we achieved promising results obtaining a sensitivity of 0.87 and a specificity of 0.65. In the future, We plan to explore more powerful features and sophisticated model architectures to increase the robustness of the predictive model to identify fatigue from speech.

7. References

[1] J. Ware, M. Kosinski, and B. Gandek, "SF-36 Health Survey: Manual & Interpretation Guide," *Lincoln, RI: QualityMetric Incorporated*, Jan. 1993.

[2] X. Gao, K. Ma, H. Yang, K. Wang, B. Fu, Y. Zhu, X. She, and B. Cui, "A rapid, non-invasive method for fatigue detection based on voice information," *Frontiers in Cell and Developmental Biology*, vol. 10, p. 994001, Sep. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9513181/

[3] J. Krajewski, U. Trutschel, M. Golz, D. Sommer, and D. Edwards, "Estimating fatigue from predetermined speech samples transmitted by operator communication systems," in *Proceedings of the 5th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : Driving Assessment 2009*. Big Sky, Montana, USA>: University of Iowa, 2009, pp. 468–474. [Online]. Available: http://pubs.lib.uiowa.edu/driving/article/id/28109/

[4] A. M. Rashwan, M. S. Kamel, and F. Karray, "Car driver fatigue monitoring using Hidden Markov Models and Bayesian networks," in *Proceedings of the International Conference on Connected Vehicles and Expo (ICCVE)*, Dec. 2013, pp. 247–251, iSSN: 2378-1297.

[5] C. Craye, A. Rashwan, M. S. Kamel, and F. Karray, "A multimodal driver fatigue and distraction assessment system," *International Journal of Intelligent Transportation Systems Research*, vol. 14, Mar. 2015.

[6] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. Ji, T. Koike, X. Li, Z. Zhang, Y. Yamamoto, and B. W. Schuller, "An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety," May 2020, arXiv:2005.00096 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2005.00096

[7] A. König, K. Riviere, N. Linz, H. Lindsay, J. Elbaum, R. Fabre, A. Derreumaux, and P. Robert, "Measuring stress in health professionals over the phone using automatic speech

analysis during the COVID-19 pandemic: Observational pilot study," *Journal of Medical Internet Research*, vol. 23, no. 4, p. e24191, Apr. 2021. [Online]. Available: https://www.jmir.org/2021/4/e24191

[8] T. Kamitani, Y. Yamamoto, N. Kurita, S. Yamazaki, S. Fukuma, K. Otani, M. Sekiguchi, Y. Onishi, M. Takegami, R. Ono, S. Konno, S. Kikuchi, and S. Fukuhara, "Longitudinal association between subjective fatigue and future falls in community-dwelling older adults: The Locomotive Syndrome and Health Outcomes in the Aizu Cohort Study (LOHAS)," *Journal of Aging and Health*, vol. 31, no. 1, pp. 67–84, Jan. 2019.

[9] C. L. Wimmelmann, N. K. Andersen, M. S. Grønkjaer, E. R. Hegelund, and T. Flensborg-Madsen, "Satisfaction with life and SF-36 vitality predict risk of ischemic heart disease: a prospective cohort study," *Scandinavian cardiovascular journal: SCJ*, vol. 55, no. 3, pp. 138–144, Jun. 2021.

[10] N. K. Andersen, C. L. Wimmelmann, E. L. Mortensen, and T. Flensborg-Madsen, "Longitudinal associations of self-reported satisfaction with life and vitality with risk of mortality," *Journal of Psychosomatic Research*, vol. 147, p. 110529, Aug. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022399921001744

[11] V. Burholt and P. Nash, "Short Form 36 (SF-36) Health Survey Questionnaire: normative data for Wales," *Journal of Public Health*, vol. 33, no. 4, pp. 587–603, Dec. 2011. [Online]. Available: https://doi.org/10.1093/pubmed/fdr006

[12] L. Lins and F. M. Carvalho, "SF-36 total score as a single measure of health-related quality of life: Scoping review," *SAGE Open Medicine*, vol. 4, p. 2050312116671725, Oct. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5052926/

[13] V. Berisha, C. Krantsevich, G. Stegmann, S. Hahn, and J. Liss, "Are reported accuracies in the clinical speech machine learning literature overoptimistic?" in *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2022, pp. 2453–2457.

[14] A. Pinto-Meza, A. Serrano-Blanco, M. T. Peñarrubia, E. Blanco, and J. M. Haro, "Assessing depression in primary care with the phq-9: can it be carried out over the telephone?" *Journal of General Internal Medicine*, vol. 20, no. 8, pp. 738–742, Aug 2005.

[15] Y. Conwell, A. Simning, N. Driffill, Y. Xia, X. Tu, S. P. Messing, and D. Oslin, "Validation of telephone-based behavioral assessments in aging services clients," *International Psychogeriatrics*, vol. 30, no. 1, pp. 95–102, Jan 2018.

[16] S. Fukuhara, S. Bito, J. Green, A. Hsiao, and K. Kurokawa, "Translation, adaptation, and validation of the SF-36 Health Survey for use in Japan," *Journal of clinical epidemiology*, vol. 51, no. 11, pp. 1037–1044, 1998.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

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

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[20] A. Khosravani, C. Glackin, N. Dugan, G. Chollet, and N. Cannings, "The Intelligent Voice 2016 Speaker Recognition System," *CoRR*, vol. abs/1611.00514, 2016. [Online]. Available: http://arxiv.org/abs/1611.00514

[21] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," in *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo, Japan: ISCA & IEEE, 2003.

[22] A. Canavan and G. Zipperlen, "CALLHOME Japanese Speech," *LDC96S37. Web Download*, 1996.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[24] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 517–520 vol.1.

[25] 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 *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[26] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[28] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.

[29] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The Interspeech 2016 Computational Paralinguistics Challenge: Deception, sincerity & native language," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech), Vols 1-5*, 2016, pp. 2001–2005.

[30] B. Schuller, S. Steidl, A. Batliner, E. Noeth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The Interspeech 2012 Speaker Trait Challenge," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech)*, vol. 1, 09 2012.

[31] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge," *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.

[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[33] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.