Towards interpretable speech biomarkers: explaining MFCC2

Brian Tracey (brian.tracey@takeda.com)
Takeda Pharmaceuticals

Dmitri Volfson
Takeda Pharmaceuticals

James Glass
Massachusetts Institute of Technology

R'mani Haulcy
Massachusetts Institute of Technology

Melissa Kostrzebski
University of Rochester

Jamie Adams
University of Rochester

Tairmae Kangarloo
Takeda Pharmaceuticals

Amy Brodtmann
Monash University

E. Dorsey
University of Rochester Medical Center

Adam Vogel
University of Melbourne

Article

Keywords:

Posted Date: February 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2478289/v1

Additional Declarations: Competing interest reported. B. Tracey, D. Volfson and T. Kangarloo are full-time employees of and own stock in Takeda Pharmaceuticals. M. Kostrzebski holds stock in Apple, Inc. J.

Adams has received compensation for consulting services from VisualDx and the Huntington Study Group; and research support from Biogen, Biosensics, Huntington Study Group, Michael J. Fox Foundation, National Institutes of Health/National Institute of Neurological Disorders and Stroke, NeuroNext Network, and Safra Foundation. E. Dorsey has received compensation for consulting services from Abbott, Abbvie, Acadia, Acorda, Bial-Biotech Investments, Inc., Biogen, Boehringer Ingelheim, California Pacific Medical Center, Caraway Therapeutics, Curasen Therapeutics, Denali Therapeutics, Eli Lilly, Genentech/Roche, Grand Rounds, Huntington Study Group, Informa Pharma Consulting, Karger Publications, LifeSciences Consultants, MCM Education, Mediflix, Medopad, Medrhythms, Merck, Michael J. Fox Foundation, NACCME, Neurocrine, NeuroDerm, NIH, Novartis, Origent Data Sciences, Otsuka, Physician's Education Resource, Praxis, PRIME Education, Roach, Brown, McCarthy and Gruber, Sanofi, Seminal Healthcare, Spark, Springer Healthcare, Sunovion Pharma, Theravance, Voyager and WebMD; research support from Biosensics, Burroughs Wellcome Fund, CuraSen, Greater Rochester Health Foundation, Huntington Study Group, Michael J. Fox Foundation, National Institutes of Health, Patient-Centered Outcomes Research Institute, Pfizer, PhotoPharmics, Safra Foundation, and Wave Life Sciences; editorial services for Karger Publications; stock in Included Health and in Mediflix, and ownership interests in SemCap. J. Glass and R. Haulcy have received research funding from Takeda Pharmaceuticals. A.Vogel is Chief Science Officer of Redenlab P/L who undertake speech biomarker research services. A. Brodtmann has no competing financial interests.

Version of Record: A version of this preprint was published at Scientific Reports on December 21st, 2023. See the published version at https://doi.org/10.1038/s41598-023-49352-2.

Towards interpretable speech biomarkers: explaining MFCC2

Brian Tracey[1,*], Dmitri Volfson[1], James Glass[3], R'mani Haulcy[3], Melissa Kostrzebski[2], Jamie Adams[2], Tairmae Kangarloo[1], Amy Brodtmann[4,5], E. Ray Dorsey[2], and Adam Vogel[5,6]

[1]Takeda Pharamaceuticals, Data Science Institute, Cambridge, MA, 02142, USA
[2]Center for Health + Technology (CHeT), University of Rochester Medical Center, Rochester, NY, USA.
[3]Massachusetts Institute of Technology, CSAIL, Cambridge, MA , 02139, USA
[4]Monash University, Melbourne, Victoria, Australia
[5]University of Melbourne, Parkville, Victoria, 3010, Australia
[6]Redenlab Inc, Melbourne, Victoria, 3010, Australia
[*]brian.tracey@takeda.com

ABSTRACT

Speech biomarkers of disease have attracted increased clinical interest in recent years, but interpretation of clinical features derived from signal processing or machine learning approaches remains challenging. As an example, the second Mel frequency cepstral coefficient (MFCC2) has been identified in several studies as a useful marker of disease, but continues to be treated as uninterpretable. Here we show that MFCC2 can be interpreted as a weighted ratio of low- to high-frequency energy, a concept which has been previously linked to disease-induced increases in aspiration noise caused by incomplete vocal fold closure, and also show how its sensitivity to disease can be increased by adjusting computation parameters.

Introduction

The last decade has seen an increase in the use of speech for health monitoring, with a focus on studies in neurological[1,2] and respiratory disease[3,4]. This is in part driven by increased ease in recording good-quality data using either smartphones[5] or cloud-based platforms[6]. Analysis of this data has used a mix of interpretable endpoints (prosodic measures related to timing and pitch, etc.) as well as speech parameterizations originally developed for speech recognition. This later category of parameterizations (which includes MFCCs, or Mel Frequency Cepstral coefficients[1], RASTA coefficients, and deep learning-derived embeddings[7]) often leads to high performance in classification or regression tasks, but low interpretability. This lack of interpretability makes it difficult to link acoustic features to disease biology and diminishes their utility to clinicians and patients.

Smartphone-based measurements can be useful in tracking Parkinson's Disease (PD)[8]. In Roche's PD Mobile Application designed for generating exploratory outcome measures for clinical trials in Parkinson's disease, speech was evaluated via a sustained vowel phonation task ("say aaah"), leading to a finding that the second MFCC coefficient (MFCC2) separated PD participants from healthy controls. This work cited earlier studies[9,10] supporting the use of MFCC2 as a basis for its inclusion as the only published speech feature in the study.

Motivated by this and other work reporting the value of MFCC2[11,12] , we examined this feature in several datasets (described below) and observed that it appears to be a useful feature in both PD and frontotemporal dementia (FTD). Given that the apparent sensitivity of MFCC2 to disease, we seek to build a better understanding of the utility and interpretability of this potentially important feature.

In what follows, we first demonstrate that MFCC2 has a relatively straightforward interpretation as a weighted ratio of low- to high-frequency energy, which has been previously identified as being linked to voice pathology. We then examine MFCC2 in three datasets, showing that a) by tuning MFCC2 calculation to include more high frequencies, we can improve its performance, and b) MFCC2 appears to depend strongly on sex but not age. We also examine MFCC2 in several datasets, showing that trends are similar across datasets but values may depend on recording quality. Further work is needed to explore this possible dependence.

MFCC2 interpretation Figure 1A) briefly summarizes the MFCC calculation. The input signal is first transformed to create a spectrogram. Mel frequency filters are then applied to resample the frequency axis in a manner that mimics the roughly logarithmic pitch sensitivity of human hearing, with finer resolution at lower frequencies and coarser resolution at high frequencies. The Mel-filtered data are then log-transformed and processed with a cepstral transform, which amounts to multiplying the Mel spectra by a series of cosine terms. As shown in Figure 1A), MFCC1 is a constant feature capturing overall energy, MFCC2 is a half-cycle of cosine, etc. The MFCC coefficients are computed by multiplying the log-transformed Mel spectra by the cosine terms and then summing across frequency.

Figure 1B) shows the cosine term associated with MFCC2, remapped from Mel frequency back to actual frequency in Hz. This figure suggests that MFCC2 is adding a weighted sum of low frequency log(energy) and subtracting off a weighted sum of high frequency log(energy). As $log(a) - log(b) = log(a/b)$, MFCC2 can be interpreted as a form of low-to-high frequency energy ratio, with the lowest and highest frequencies contributing most strongly due to the weighting applied.

Figure 1B) also shows two MFCC2 curves. MFCC2 computation requires the user to specify the maximum frequency used in calculation. While 8 kHz is a common upper limit, we show below that increasing the maximum frequency can be beneficial.

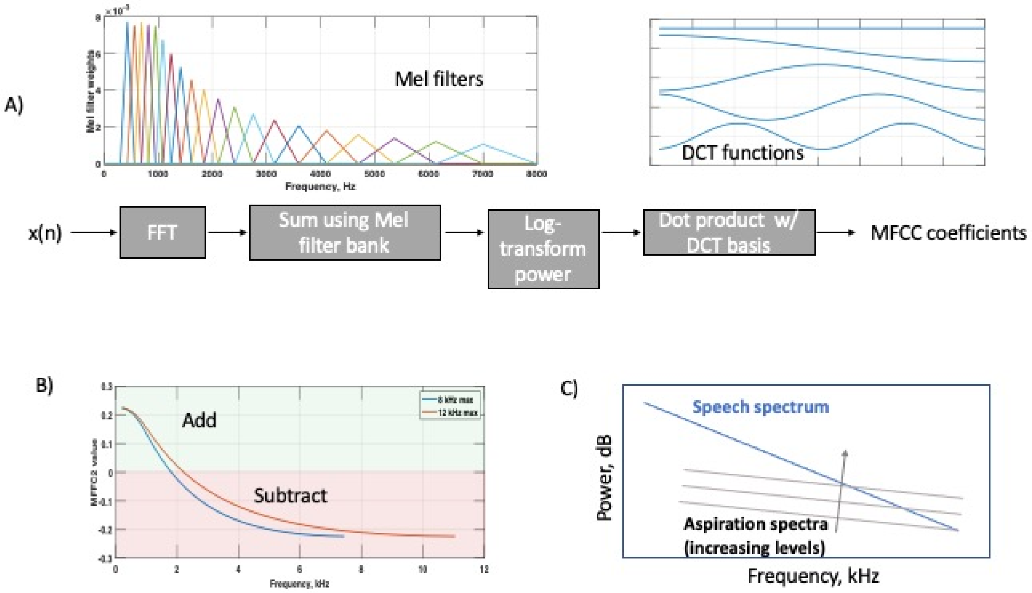


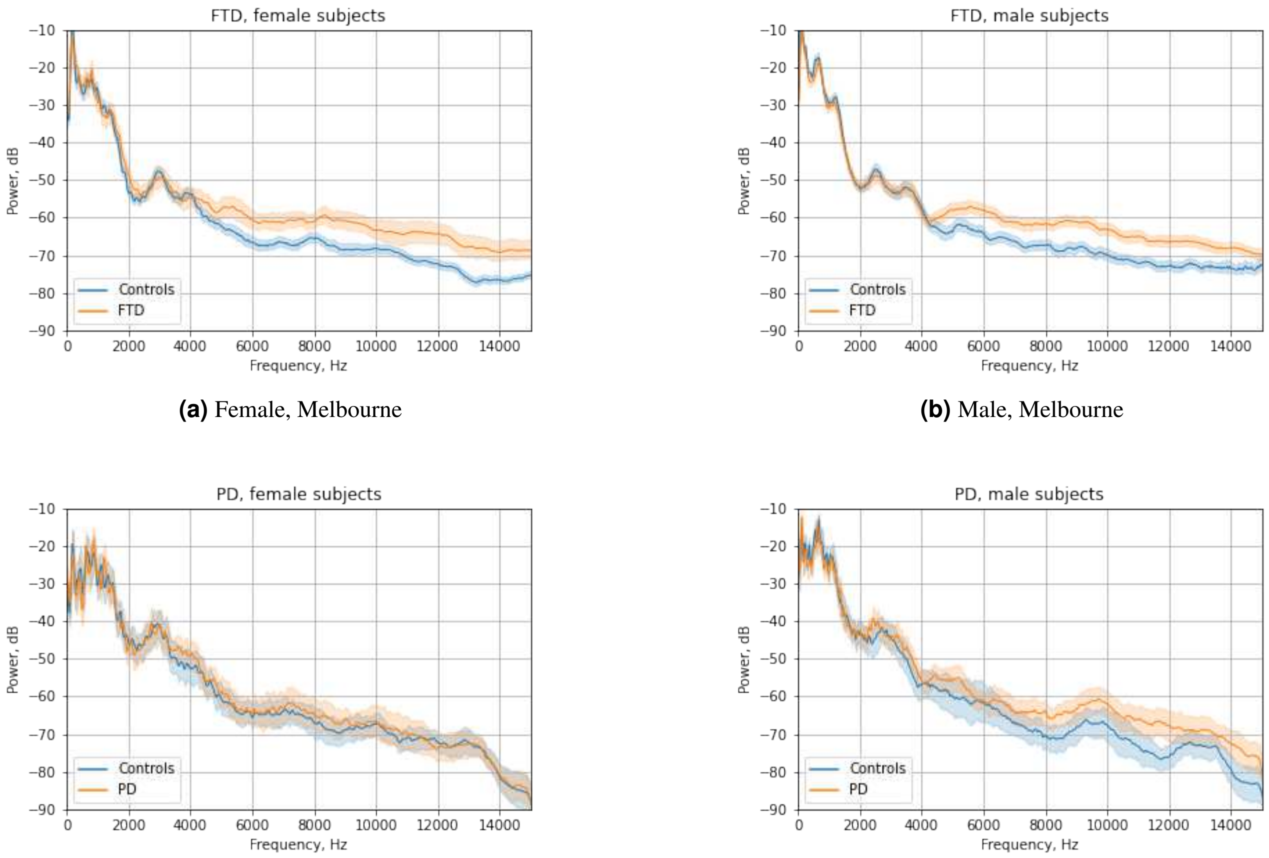
Figure 1. Overview of MFCC calculation; a) shows a schematic of the MFCC computation; b) shows the MFCC2 cosine term mapped to frequency in Hz, for maximum frequency values of 8 and 12 kHz; c) shows a 'cartoon' view illustrating how increasing aspiration noise can impact affect the overall spectrum at high frequencies.

Given that MFCC2 can be interpreted as a low-to-high energy ratio, the next step is to relate this to voice quality and/or disease. Breathy voice can be characterized in the low-frequency range via increased amplitude of the first harmonic, as the glottal waveform becomes more rounded due to non-simultaneous closure along the length of the vocal cords[13]. More relevant to MFCC2, high-frequency energy also increases due to the presence of turbulent airflow associated with breathy voice. Hillenbrand and Houde note[13], "... the presence of aspiration noise, which is stronger in the mid and high frequencies than in the lows, can result in a voice signal that is richer in high-frequency energy than nonbreathy signals." This concept is illustrated in a 'cartoon' view in Figure 1C); in the absence of aspiration noise, voice signal levels typically fall off rapidly as frequency increases. Aspiration noise, generated by turbulent flow, tends to fall off less rapidly with frequency. As aspiration noise increases, it primarily impacts the high-frequency part of the overall (voice plus aspiration noise) spectrum but remains negligible compared to voice at the low frequencies.

Hillenbrand and Houde proposed a high-to-low (H/L) ratio, comparing average energy above 4 kHz to average energy below 4 kHz, to capture aspiration noise. Thus it is anticipated that voices with more aspiration noise, caused by incomplete vocal fold closure, will have lower MFCC2 values and higher H/L ratios. Similarly, ratios splitting high and low frequencies appear sensitive to changes in speech and voice resulting from Huntington's disease[14], illicit drug use[15], congestion[16], and fatigue [17].

Results

We examined sustained phonation recordings from three datasets, with patient characteristics shown in Table 1. All three datasets contain control participants. The University of Melbourne/Monash dataset also includes participants with FTD (with



(a) Female, Melbourne

(b) Male, Melbourne

(c) Female, WatchPD

(d) Male, WatchPD

Figure 2. Mean spectra (with 95th percentile confidence limits) for Melbourne and WatchPD datasets. Note the generally higher values above 4 kHz for non-healthy participants.

behavioral variant FTD most common), while the WatchPD dataset includes participants with PD. The PD participants were recruited soon after diagnosis so have fewer years of disease than the FTD participants.

Dataset	Diagnosis	Age	Years since Diagnosis	# Participants (Male/Female/Other)
Melbourne	FTD	65.0 (60.2 / 71.0)	3 (2 / 5.5)	36 / 19 / 0
Melbourne	Controls	63.0 (56.0 / 70.0)	-	50 / 56 / 0
WatchPD	PD	66.0 (55.0 / 71.0)	<1	29 / 25 / 0
WatchPD	Controls	61.0 (54.5 / 69.5)	-	19 / 24 / 0
CLAC	Controls	33.0 (27.0 / 42.0)	-	799 / 800 / 11

Table 1. Subject information. Age and Years since Diagnosis are listed as median (25th percentile / 75th percentile).

Figure 2 shows the averaged acoustic spectra for the Melbourne and WatchPD datasets, comparing controls to participants with neurological disease (FTD or PD) (called "cases" below). Because spectral characteristics vary by sex, plots are shown separately for males and females. In general, these plots indicate that FTD/PD participants have higher acoustic power at high frequencies as compared to controls. This is seemingly more evident in men, as well as in the FTD participants, who have greater years of disease duration.

We next computed various metrics to capture low-to-high frequency energy ratios. We computed MFCC2 with the standard 8 kHz maximum frequency, and MFCC2 with a 12 kHz maximum frequency. Like Hillenbrand and Houde, we compare energy above and below 4 kHz using an Energy Ratio metric; while they computed a high-to-low ratio, we instead compute a low-to-high ratio for easier comparison to MFCC2 (see Methods). Table 2 lists AUC (area under the curve) values for these metrics from ROC curves for Melbourne and WatchPD datasets (no ROC curves are shown for CLAC as the database only

Metric	Dataset	AUC, Female	AUC, Male
MFCC2, 12kHz	Melbourne	**0.82 (0.66-0.97)**	**0.78 (0.67-0.89)**
MFCC2, 8kHz	Melbourne	0.69 (0.51-0.86)	0.73 (0.61-0.86)
Energy Ratio	Melbourne	0.73 (0.58-0.87)	0.75 (0.64-0.87)
MFCC2, 12kHz	WatchPD	**0.61 (0.45-0.77)**	**0.69 (0.54-0.84)**
MFCC2, 8kHz	WatchPD	**0.61 (0.45-0.77)**	0.67 (0.51-0.83)
Energy Ratio	WatchPD	0.59 (0.43-0.75)	0.58 (0.41-0.75)

Table 2. ROC Area under the curve (AUC) for different metrics and datasets; mean AUC and 95th percentile confidence intervals are shown. Bold values highlight the best-performing metric for each dataset.

			MFCC2, 12kHz		MFCC2, 8kHz		Energy Ratio	
Dataset	Gender	Diagnosis	M	SD	M	SD	M	SD
WatchPD	female	Healthy	136.0	22.3	95.8	19.1	33.5	6.8
WatchPD	female	PD	128.1	21.3	89.6	16.0	31.1	6.8
WatchPD	male	Healthy	148.8	20.6	107.4	15.7	34.9	8.1
WatchPD	male	PD	133.2	20.4	98.3	14.5	32.3	6.6
Melbourne	female	Healthy	143.4	11.8	109.9	10.7	38.3	4.7
Melbourne	female	FTD	126.8	18.8	101.5	13.0	34.5	4.1
Melbourne	male	Healthy	148.4	13.1	112.5	8.5	38.8	4.1
Melbourne	male	FTD	131.9	14.9	103.5	10.4	34.7	4.5
CLAC	female	Healthy	131.1	29.2	90.8	24.4	30.1	9.1
CLAC	male	Healthy	137.1	29.9	96.0	23.8	30.4	9.8
CLAC	other	Healthy	123.2	30.6	85.1	30.2	27.6	9.9

Table 3. Descriptive statistics for MFCC2-related metrics, by dataset, sex and diagnosis.

contained control participants). MFCC2 with 12 kHz maximum frequency appears to show best separability, with the energy ratio being least discriminative. Descriptive statistics for these metrics are shown in Table 3. In the Melbourne dataset, a small number of participants made more than one clinic visit, so Tables 2-3 include only the first visit for each subject. Note that in all cases, control participants have higher mean values for all three metrics.

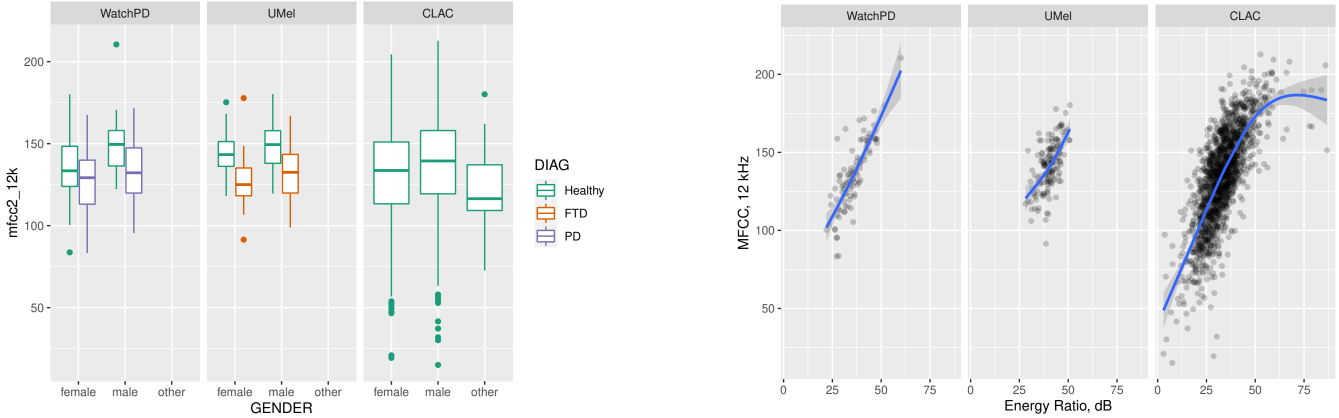
Statistical analysis: We next explored whether these metrics were dependent on sex, age and dataset, as well as diagnosis. We modeled each MFCC2 endpoint as a linear combination of factors. A model selection process using the Akaike Information Criterion (AIC) led to selection of a model which includes gender, dataset, and age. Because dataset and diagnosis are confounded, this analysis was done for control participants only. For MFCC2 with fmax = 12 kHz, there was a highly significant effect of gender (males were higher, $p<0.001$) and a significant effect of dataset (CLAC values were lower, $p <0.05$) with no significant effect of age. The corresponding boxplots for MFCC2 with fmax = 12 kHz are shown in Figure 3a) for the different datasets, by gender and diagnosis. MFCC2 values are higher in men, reflecting the increased low-frequency content in these speakers (and are lower in disease, likely reflecting the aspiration noise discussed above). For MFCC2 with fmax = 8 kHz (not plotted), these findings were repeated, but also there were also significant effects of age (values decreased with a small slope of roughly 1 point per decade, $p<0.05$) and also Melbourne values were significantly higher than WatchPD values.

Figure 3b) shows the relationship between MFCC2 and the Energy Ratio metric described above, with moderate to good correlations (Pearson correlations are 0.82 in the WatchPD data, 0.60 in the Melbourne data, and 0.74 in CLAC). Note that the CLAC outliers observed may be due to noise cancellation artifacts, as discussed below.

Discussion

The results suggest that MFCC2 is capturing a tendency toward increased high-frequency acoustic energy in phonation of individuals with neurological disease. This tendency may reflect impaired motor control and thus and thus incomplete closure of the vocal folds.

The acoustic spectra in Figure 2 suggest that differences between cases and controls increase with frequency, which is consistent with broadband aspiration noise. Thus, increasing the maximum frequency for MFCC2 from 8 kHz to 12 kHz



(a) MFCC2 dependence on demographics

(b) Correlation with Energy Ratio

Figure 3. MFCC2 (fmax=12 kHz) characteristics: a) dependence on gender and diagnosis, and b) correlation with 4 kHz Energy Ratio metric, by dataset, with loess fit. Detailed analysis found in the Statistical Analysis section.

appears to improve the ability of MFCC2 to separate cases from controls. In principle, further increasing the upper range could further improve separability between groups. However, manual examination of our data shows occasional high-frequency noise of undetermined source at frequencies above roughly 12-15 kHz. Thus, we limited the upper frequency range to 12 kHz in our calculation. This upper limit is based on engineering judgement, not extensive data exploration. If phonations were recorded in a very quiet environment using high quality hardware[18], there could be benefit to including higher frequencies.

Because MFCC2 values depend on the frequency limits used in computation, it is important that MFCC parameter settings should be reported along with findings, to better allow comparisons between different studies. Widely used MFCC implementations such as librosa[19] default to using half the sampling rate as the upper limit, which means that MFCC values could easily vary depending on recording settings (and may include very high-frequency noise, as noted above).

As noted above, there is a clear impact of gender on MFCC2. In healthy participants, MFCC2 values are higher in men than women in all cohorts (Table 3), presumably reflecting the fact that male voices are skewed to lower frequencies. Interestingly, Table 2 shows that MFCC2 better discriminates Parkinson's in men (higher AUC). It has been previously established that PD differentially affects male and female voices[1,7]. A possible physical explanation is that male voices containing relatively less mid-to-high frequency energy would be more affected by addition of aspiration noise in this frequency range (see Fig. 1C). However, the same pattern is not observed in FTD participants. Whether this is a true physiological difference, or is related to the relatively small number of women in the FTD cohort, is unclear. In both cases the uncertainty in the AUCs (seen in the confidence bounds) argues for repeating these analyses in additional datasets.

Spectral differences and MFCC2 differences between cases and controls were clearer in the Melbourne dataset (FTD) than the WatchPD dataset (PD). This may partly reflect the advanced disease state of the FTD cohort (longer disease duration) than the early-stage PD participants in WatchPD, but may also be impacted by recording quality. Fully understanding the progression of the disease in FTD vs. PD would require datasets with a larger range of progression or a longitudinal sample.

There are several potential reasons for the observed dependence on dataset. The Melbourne dataset consists of Australian speakers, while the WatchPD and CLAC cohorts are US-based. Perhaps more importantly, the datasets use different recording setups; Melbourne is a mixture of in-lab recordings and at-home (smartphone) recordings, WatchPD consists of in-clinic recordings made with an iPhone [1], and CLAC was recorded via internet browsers using Amazon Mechanical Turk. This means CLAC data are subject to data compression artifacts that may be variable depending on browser, service provider, etc. Manual review uncovered CLAC recordings in which the sustained phonation was distorted after the first second or so, perhaps because noise cancellation algorithms incorrectly identify the sustained phonation as a form of background noise (this issue impacts sustained phonation more than regular speech, as algorithms presumably target hum-like signals). These variations in browser processing may explain the noticeably higher standard deviations seen in metrics computed from CLAC (Table 3). This suggests that the acoustic quality of internet-acquired data be carefully reviewed, and that ideally browser settings be controlled to disable processing algorithms, especially when subtle acoustic features of speech are being analyzed.

Since MFCC2 was not optimized for disease detection, it is likely possible to find weighted energy ratios that would discriminate cases from controls better than MFCC2. Ideally, such features could be tuned separately by gender. Such exploration would benefit from larger datasets to ensure generalizability of the results.

[1] note that only data from the WatchPD baseline visit are analyzed here; the complete WatchPD dataset also includes at-home recordings

Finally, while we focused on MFCC2, we note that several papers report higher-order MFCC coefficients as being useful features[1]. Just as MFCC2 is interpreted as a low-to-high energy comparison, it would be possible to interpret MFCC3 as a mid-frequency to (low+high)-frequency ratio, MFCC4 as a ratio of two frequency bands to two slightly higher frequency bands, etc. (see the cosine shapes in Fig. 1a). However, the physical / biological meaning of such ratios becomes increasingly unclear as the MFCC coefficient number increases. Features such as spectral contrast or spectral flatness[19] may give more intuitive ways to probe finer-scale acoustic structure. The time-derivatives of higher MFCC coefficients (often computed as a speech recognition feature) do have some degree of interpretability as measures of the temporal stability of the acoustic spectrum.

Methods

We analyzed recordings of sustained vowel phonation ("aaah") from baseline clinic visits in the WatchPD study[5], and from data collected at the University of Melbourne (consisting of healthy elderly controls[20] and FTD participants[2]). We also utilize the public-domain CLAC dataset[6] of normative speakers, collected using Amazon Mechanical Turk.

In each dataset, recordings were first automatically segmented using custom Python code to identify the vowel phonation. Processing first detected voiced regions of speech using the voicing/pitch detection from Parselmouth[21]. The initial 0.75 sec of voiced data were discarded to remove transient effects, and the next 2.5 sec were retained for analysis. If no segment was detected, the recording was not analyzed further. The segmented waveform was then processed using Librosa[19] to compute MFCC coefficients, using 133 Hz as a lower bound and either 8 or 12 kHz as upper bounds. In addition, the acoustic spectra were computed using the scipy-signal implementation of the Welch periodogram method, using 20 msec, 50% overlapped Hanning windows, as plotted in Fig. 2. These (linear) spectra $P_{xx}(f)$ were used to compute the Energy Ratio metric:

$$ER, dB = 10\log_{10}\left(\frac{\int_0^{4000} P_{xx}(f)df}{\int_{4000}^{f_{max}} P_{xx}(f)df}\right) \tag{1}$$

where $f_{max} = 12,000$ Hz. Note that whereas Hillenbrand and Houde[13] formed a ratio of high-to-low energy, we compute low-to-high for easier comparison with MFCC2. Statistical analysis was performed in R version 4.1.0. AUC analysis was performed using the pROC package (version 1.18.0) which uses bootstrapping to estimate confidence intervals.

Data availability

Raw audio for the CLAC dataset is available at https://groups.csail.mit.edu/sls/downloads/clac/. The extracted features for the CLAC dataset are available at https://github.com/brianhtracey/mfcc2_related. Extracted features for other datasets may be available upon reasonable request by contacting Brian Tracey (brian.tracey@takeda.com).

Code availability

Core python code for feature extraction is available at https://github.com/brianhtracey/mfcc2_related.

Acknowledgements (not compulsory)

This research was supported by Millennium Pharmaceuticals, Inc. (a subsidiary of Takeda Pharmaceuticals). Funding for the WatchPD study was contributed by Biogen, Takeda, and the members of the Critical Path for Parkinson's Consortium 3DT Initiative, Stage 2. The authors acknowledge the members and advisors of the CPP 3DT initiative and the FDA for their insightful feedback to the WATCH-PD study design and execution. The authors gratefully acknowledge WatchPD participants who joined WatchPD study participant forums, which helped clarify the need for more interpretable voice features. We also acknowledge Dr. Sandra Rojas as contributor of healthy control data from Australia.

Author contributions statement

B.T., D.V., and J.G. conceived the experiment(s), M.K.,J.A.,R.D.,R.H.,J.G. T.K., and A.B. conducted the experiment(s) and contributed participants, B.T. and D.V. analysed the results. All authors reviewed the manuscript.

Additional information

B. Tracey, D. Volfson and T. Kangarloo are full-time employees of and own stock in Takeda Pharmaceuticals. M. Kostrzebski holds stock in Apple, Inc. J. Adams has received compensation for consulting services from VisualDx and the Huntington Study Group; and research support from Biogen, Biosensics, Huntington Study Group, Michael J. Fox Foundation, National Institutes of Health/National Institute of Neurological Disorders and Stroke, NeuroNext Network, and Safra Foundation. E. Dorsey has received compensation for consulting services from Abbott, Abbvie, Acadia, Acorda, Bial-Biotech Investments, Inc., Biogen, Boehringer Ingelheim, California Pacific Medical Center, Caraway Therapeutics, Curasen Therapeutics, Denali Therapeutics, Eli Lilly, Genentech/Roche, Grand Rounds, Huntington Study Group, Informa Pharma Consulting, Karger Publications, LifeSciences Consultants, MCM Education, Mediflix, Medopad, Medrhythms, Merck, Michael J. Fox Foundation, NACCME, Neurocrine, NeuroDerm, NIH, Novartis, Origent Data Sciences, Otsuka, Physician's Education Resource, Praxis, PRIME Education, Roach, Brown, McCarthy and Gruber, Sanofi, Seminal Healthcare, Spark, Springer Healthcare, Sunovion Pharma, Theravance, Voyager and WebMD; research support from Biosensics, Burroughs Wellcome Fund, CuraSen, Greater Rochester Health Foundation, Huntington Study Group, Michael J. Fox Foundation, National Institutes of Health, Patient-Centered Outcomes Research Institute, Pfizer, PhotoPharmics, Safra Foundation, and Wave Life Sciences; editorial services for Karger Publications; stock in Included Health and in Mediflix, and ownership interests in SemCap. J. Glass and R. Haulcy have received research funding from Takeda Pharmaceuticals. A.Vogel is Chief Science Officer of Redenlab P/L who undertake speech biomarker research services. A. Brodtmann has no competing financial interests.

References

1. Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J. & Ramig, L. O. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE transactions on biomedical engineering* **59**, 1264–1271 (2012).
2. Vogel, A. P. *et al.* Motor speech signature of behavioral variant frontotemporal dementia: Refining the phenotype. *Neurology* **89**, 837–844 (2017).
3. Quatieri, T. F., Talkar, T. & Palmer, J. S. A framework for biomarkers of covid-19 based on coordination of speech-production subsystems. *IEEE Open J. Eng. Medicine Biol.* **1**, 203–206 (2020).
4. Tracey, B. *et al.* Voice biomarkers of recovery from acute respiratory illness. *IEEE J. Biomed. Heal. Informatics* (2021).
5. Cedarbaum, J. M. *et al.* Enabling efficient use of digital health technologies to support parkinson's disease drug development through precompetitive collaboration. In *American Society for Clinical Pharmacology & Therapeutics (ASCPT) Meeting* (2019).
6. Haulcy, R. & Glass, J. CLAC: A Speech Corpus of Healthy English Speakers. In *Proc. Interspeech 2021*, 2966–2970, DOI: 10.21437/Interspeech.2021-1810 (2021).
7. Jeancolas, L. *et al.* X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech. *Front. Neuroinformatics* **15**, 578369 (2021).
8. Lipsmeier, F. *et al.* Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 parkinson's disease clinical trial. *Mov. Disord.* **33**, 1287–1297 (2018).
9. Kapoor, T. & Sharma, R. Parkinson's disease diagnosis using mel-frequency cepstral coefficients and vector quantization. *Int. J. Comput. Appl.* **14**, 43–46 (2011).
10. Benba, A., Jilbab, A. & Hammouch, A. Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines. *Int. J. on Electr. Eng. Informatics* **7**, 297 (2015).
11. Taguchi, T. *et al.* Major depressive disorder discrimination using vocal acoustic features. *J. affective disorders* **225**, 214–220 (2018).
12. Al-Hameed, S., Benaissa, M. & Christensen, H. Simple and robust audio-based detection of biomarkers for alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 32–36 (2016).
13. Hillenbrand, J. & Houde, R. A. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *J. Speech, Lang. Hear. Res.* **39**, 311–321 (1996).
14. Vogel, A. P., Shirbin, C., Churchyard, A. J. & Stout, J. C. Speech acoustic markers of early stage and prodromal huntington's disease: a marker of disease onset? *Neuropsychologia* **50**, 3273–3278 (2012).
15. Vogel, A. P. *et al.* Adults with a history of recreational cannabis use have altered speech production. *Drug Alcohol Dependence* **227**, 108963 (2021).
16. Lee, G.-S., Yang, C. C., Wang, C.-P. & Kuo, T. B. Effect of nasal decongestion on voice spectrum of a nasal consonant-vowel. *J. Voice* **19**, 71–77 (2005).
17. Vogel, A. P., Fletcher, J. & Maruff, P. Acoustic analysis of the effects of sustained wakefulness on speech. *The J. Acoust. Soc. Am.* **128**, 3747–3756 (2010).

18. Vogel, A. P. & Reece, H. Recording speech: Methods and formats. In *Manual of Clinical Phonetics*, 217–227 (Routledge, 2021).

19. McFee, B. *et al.* librosa 0.5.0, DOI: 10.5281/zenodo.293021 (2017).

20. Schultz, B. G., Rojas, S., St John, M., Kefalianos, E. & Vogel, A. P. A cross-sectional study of perceptual and acoustic voice characteristics in healthy aging. *J. Voice* (2021).

21. Jadoul, Y., Thompson, B. & de Boer, B. Introducing parselmouth: A python interface to praat. *J. Phonetics* **71**, 1–15, DOI: doi.org/10.1016/j.wocn.2018.07.001 (2018).