Journal of Affective Disorders 225 (2018) 214–220

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

# Major depressive disorder discrimination using vocal acoustic features

Takaya Taguchi[a,b], Hirokazu Tachikawa[a,c,*], Kiyotaka Nemoto[a,c], Masayuki Suzuki[d], Toru Nagano[d], Ryuki Tachibana[d], Masafumi Nishimura[e], Tetsuaki Arai[a,c]

[a] *Department of Psychiatry, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan*
[b] *University of Tsukuba Hospital, Japan*
[c] *Department of Psychiatry, Faculty of Medicine, University of Tsukuba, Japan*
[d] *IBM Japan, LTD., IBM Research, Tokyo, Japan*
[e] *Graduate School of Integrated Science and Technology, Shizuoka University, Japan*

## ARTICLE INFO

*Keywords:*
Voice analysis
Acoustic features
Depression
Biomarker
MFCC
Discriminant analysis

## ABSTRACT

*Background:* The voice carries various information produced by vibrations of the vocal cords and the vocal tract. Though many studies have reported a relationship between vocal acoustic features and depression, including mel-frequency cepstrum coefficients (MFCCs) which applied to speech recognition, there have been few studies in which acoustic features allowed discrimination of patients with depressive disorder. Vocal acoustic features as biomarker of depression could make differential diagnosis of patients with depressive state. In order to achieve differential diagnosis of depression, in this preliminary study, we examined whether vocal acoustic features could allow discrimination between depressive patients and healthy controls.
*Methods:* Subjects were 36 patients who met the criteria for major depressive disorder and 36 healthy controls with no current or past psychiatric disorders. Voices of reading out digits before and after verbal fluency task were recorded. Voices were analyzed using OpenSMILE. The extracted acoustic features, including MFCCs, were used for group comparison and discriminant analysis between patients and controls.
*Results:* The second dimension of MFCC (MFCC 2) was significantly different between groups and allowed the discrimination between patients and controls with a sensitivity of 77.8% and a specificity of 86.1%. The difference in MFCC 2 between the two groups reflected an energy difference of frequency around 2000–3000 Hz.
*Conclusions:* The MFCC 2 was significantly different between depressive patients and controls. This feature could be a useful biomarker to detect major depressive disorder.
*Limitations:* Sample size was relatively small. Psychotropics could have a confounding effect on voice.

## 1. Introduction

The voice carries various information, including that beyond the verbal message, important for diagnosis or state evaluation of mental disorders. It is known that psychiatric symptoms are diagnosed not only from a patient's spoken communication, but also from emotions and non-verbal information such as intention, attitude, or physical state (Ladd, 1980). Clinical psychiatrists or psychologists have experiences having difficulties in understanding what patients with depressions say because of their muffled speech. The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) illustrates psychomotor retardation as “speech that is decreased in volume, inflection, amount, or variety of content, or muteness” and “it must be severe enough to be observable by others”. However, there are no objective criteria about these speeches of psychomotor retardation.

Acoustic features which assess voices include directly-relevant features how voices are heard directly (e.g., volume, speaking duration, pause duration, musical pitch (or fundamental frequency), and formants) and indirectly-relevant features (e.g., zero crossing rate, harmonics to noise ratio, and mel-frequency cepstrum coefficients). Zero crossing rate is the rate of sign-changes along a voice signal. Harmonics to noise ratio quantifies the amount of additive noise in the voice signal. Mel-frequency cepstrum coefficients (MFCCs) were introduced by Mermelstein in the 1970's (Davis and Mermelstein, 1980), which have been shown to reflect vocal tract changes (Yinghua Zhu et al., 2013) and have been widely used in speech recognition. The method for extracting MFCCs is as follows: 1) calculate Fast Fourier Transform (FFT) spectrum from frequency, 2) extract filterbank output allocated on a mel scale based upon human aural characteristics, and 3) get cepstrum coefficient from Discrete Cosine Transform (DCT). Mitrović et al. showed that the lowest MFCC (MFCC0) represents the average power of the spectrum, and the second MFCC (MFCC 1) approximates the broad

[*] Correspondence to: Department of Psychiatry, Graduate School of Comprehensive Human Sciences, University of Tsukuba, 1-1-1 Tennoudai Tsukuba, Ibaraki 305-8575, Japan.
*E-mail address:* tachikawa@md.tsukuba.ac.jp (H. Tachikawa).

http://dx.doi.org/10.1016/j.jad.2017.08.038
Received 15 January 2017; Received in revised form 1 June 2017; Accepted 14 August 2017
Available online 16 August 2017

shape of the spectrum. The higher-order coefficients, including MFCC 2, represent finer spectral details (Mitrović et al., 2010). Therefore, lower-order coefficients are hard to be affected by the voice pitch and coefficients except MFCC 0 can omit influence of volume.

Various studies have reported a relationship between parameters derived from the voice and depression (Tolkmitt and Scherer, 1986; Wittels et al., 2002). For example, Nilsonne and colleagues reported that the fundamental frequencies of the voices of patients with depression decreased and pauses between the interviewer's questions and the patients' answers lengthened (Nilsonne et al., 1988). Another study reported a correlation between the Hamilton Depression Rating Scale and speaking rate, pitch variability, and percent of pause time (Cannizzaro et al., 2004). Several studies have explored the possibility of utilizing various information gleaned from the voice as therapeutic markers. Mundt and colleagues reported that response to depression treatment was related to pitch variability, pauses while speaking, and speed of speaking (Mundt et al., 2007). They also reported a relationship between response to treatment and vocalization time, number of pauses, speaking rate, fundamental frequency, and formants in major depressive disorder patients (Mundt et al., 2012). Alpert et al. focused on prosody and reported that patients with depression spoke with reduced prosody compared to normal subjects and that treatment affected prosody (Alpert et al., 2001). However, directly-relevant acoustic features including vocalization time, number of pauses, and speaking rate make recording time long, and formants, fundamental frequency, and prosody are subject to be affected by the individual characteristics and hard to detect the differences between individuals. Moreover, these acoustic features only reflect how voice is heard.

In addition to these acoustic features, a recent development in voice analysis has enabled the use of a wider variety of indirectly-relevant acoustic features including MFCCs and zero-crossing rate. In the engineering field, the Interspeech 2009 Emotion Challenge (Schuller et al., 2009) and the Interspeech 2010 Paralinguistic Challenge (Schuller et al., 2010) revealed a relationship between various acoustic features and emotions or paralinguistic information (Akkaralaertsest and Yingthawornsuk, 2015; Cummins et al., 2015; Joshi et al., 2013; Low et al., 2011). Among these reports, a relationship between MFCCs and depression was frequently reported. However, most of the data used in these challenges were derived from normal subjects, and few studies explored the MFCCs derived from the voices of patients with depression (Akkaralaertsest and Yingthawornsuk, 2015; Cummins et al., 2011). Furthermore, most studies adopted approach which artificial intelligence discriminates depressive voice from healthy voice by multiple MFCCs and no studies investigated which MFCCs are related to depressive voice. If acoustic features were used as biomarker of depression, it could make differential diagnosis of depressive states. That means that voice of helpline call or interactive voice response system would be an asset for evaluation of depressive state or risk assessment of the person. Questionnaires have risk of dishonest responses. In that respect, assessment by voice could be an advantage. It could be also useful for people with dementia who have difficulty to answer questionnaires by themselves.

In order to achieve differential diagnosis of depression, as a first step, we investigated whether various vocal acoustic features, including acoustic features indirectly-relevant to how voice is heard, could allow discrimination between patients with depression and control subjects in this study. Based on previous reports, we hypothesized that some MFCCs could be descriminant factors, and thus investigated which voice properties affected MFCCs. Considering the possibility of clinical application, we evaluated short utterances. In addition, we compared voices before and after a verbal fluency task (VFT), which is known to activate the frontal lobe (Herrmann et al., 2003; Pu et al., 2012), to examine whether such tasks could influence acoustic features.

## 2. Methods and materials

### 2.1. Subjects

Thirty-eight patients with depression were enrolled from the department of psychiatry at the University of Tsukuba Hospital, Tsukuba University Health Center, Ibaraki Prefectural Medical Center of Psychiatry and Kurita Hospital. They all met the criteria for major depressive disorder (MDD), which were determined with clinical interviews based on the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Two patients whose voices were not recorded properly were excluded, and the remaining 36 patients were adopted for the study. The control group consisted of 66 subjects who did not have current or past psychiatric disorders and who were recruited from the local population. One subject whose voice was not recorded properly was excluded. Two participants were excluded because they had a history of major medical or neurological illness, significant head trauma, or a lifetime history of alcohol or drug dependence. In addition, we excluded 27 subjects whose scores for the Quick Inventory of Depressive Symptomatology - Self-Report, Japanese version (QIDS-SRJ) (Fujisawa, 2010; Rush et al., 2003) were higher than six (cut-off point). As a result, the subjects consisted of 36 patients with MDD (22 males and 14 females; age: 21–79 years; mean age ± standard deviation (SD): 44.0 ± 16.3), and 36 healthy control individuals (16 males and 20 females; age: 22–58 years; mean age ± SD: 38.0 ± 10.4). In depression patients, the mean duration of illness was 7.00 ± 5.73(SD) years and the median was 5.47 years. The equivalent doses (Inada and Inagaki, 2015) of psychotropic drugs (mean ± SD) were 111.0 ± 72.4 mg of imipramine in antidepressants, 54.1 ± 117.5 of chlorpromazine in antipsychotics, and 9.9 ± 11.0 of diazepam in anxiolytics. Only one patient was unmedicated. All the subjects were Japanese. Demographics for the subjects are summarized in Table 1. This study was approved by all of the ethical committees of the department of psychiatry at the University of Tsukuba Hospital, Tsukuba University Health Center, Ibaraki Prefectural Medical Center of Psychiatry, and Kurita Hospital, and written informed consent was obtained from each participant.

### 2.2. Voice recording

Each subject was asked to read out ten digits "012–345–6789" like a telephone number. Next, we administered a VFT in which subjects spoke aloud as many words as possible beginning with the vowels "a", "u", and "o" within thirty seconds. Afterwards, subjects were asked to again read out the number "012–345–6789". Each task was conducted in Japanese. Voices were recorded using a Google Nexus 7 (TM) tablet

**Table 1**
Demographics of subjects.

| | Depression patients (mean ± SD) | Controls (mean ± SD) | *p* |
|---|---|---|---|
| *n* (male/female) | 36 (22/14) | 36 (16/20) | n.s. |
| Age (years) | 44.0 ± 16.3 | 38.0 ± 10.4 | n.s. |
| QIDS-SRJ | 11.7 ± 6.2 | 2.47 ± 1.8 | < 0.001 |
| Disease duration (years) | 7.00 ± 5.73 | | |
| Antidepressants (mg) | 114.0 ± 72.4 | | |
| Antipsychotics (mg) | 54.1 ± 117.5 | | |
| Anxiolytics (mg) | 9.9 ± 11.0 | | |

SD: standard deviation.

QIDS-SRJ: Quick Inventory of Depressive Symptomatology - Self-Report, Japanese version.

n.s.: not significant.

Psychotropic drugs were categorized as antipsychotics, antidepressants and anxiolytics, and the dose of each drugs was calculated as equivalent doses (antypsychotics: chlorpromazine equivalent dose, antidepressants: imipramine equivalent dose, anxiolytics: diazepam equivalent dose) (Inada and Inagaki, 2015).

**Table 2**
Differences between Depression patients and Controls.

| Acoustic features | Depression patients (SD) | | Controls (SD) | | *p-value* | |
|---|---|---|---|---|---|---|
| | $n = 36$ | Male ($n = 22$)<br>Female ($n = 14$) | $n = 36$ | Male ($n = 16$)<br>Female ($n = 20$) | Total | Male<br>Female |
| Root mean square energy | $5.23 \times 10^{-3}$ ($3.49 \times 10^{-3}$) | $5.63 \times 10^{-3}$ ($4.01 \times 10^{-3}$) | $7.31 \times 10^{-3}$ ($7.70 \times 10^{-3}$) | $9.36 \times 10^{-3}$ ($1.06 \times 10^{-2}$) | n.s. | n.s. |
| | | $4.59 \times 10^{-3}$ ($2.49 \times 10^{-3}$) | | $5.67 \times 10^{-3}$ ($3.67 \times 10^{-3}$) | | n.s. |
| Zero crossing rate | $1.43 \times 10^{-3}$ ($1.97 \times 10^{-2}$) | $1.48 \times 10^{-3}$ ($2.13 \times 10^{-2}$) | $1.49 \times 10^{-3}$ ($1.74 \times 10^{-2}$) | $1.55 \times 10^{-3}$ ($1.78 \times 10^{-2}$) | n.s. | n.s. |
| | | $1.35 \times 10^{-3}$ ($1.44 \times 10^{-2}$) | | $1.45 \times 10^{-3}$ ($1.63 \times 10^{-2}$) | | n.s. |
| Harmonics to noise ratio | 0.351 (0.0589) | 0.323 (0.0213) | 0.370 (0.0536) | 0.341 (0.0448) | n.s. | n.s. |
| | | 0.396 (0.0673) | | 0.393 (0.0495) | | n.s. |
| Fundamental frequency | 32.54 (31.1) | 14.1 (9.62) | 35.5 (27.2) | 18.5 (16.4) | n.s. | n.s. |
| | | 61.4 (31.4) | | 49.6 (26.3) | | n.s. |
| MFCC 1 | −10.3 (1.65) | −10.3 (1.92) | −10.2 (1.94) | −9.73 (2.04) | n.s. | n.s. |
| | | −10.2 (1.16) | | −10.5 (1.83) | | n.s. |
| MFCC 2 | −10.2 (2.56) | −10.3 (2.43) | −14.3 (2.16) | −14.0 (2.18) | $< 0.001$ | $< 0.001$ |
| | | −10.1 (2.84) | | −14.5 (2.84) | | $< 0.001$ |
| MFCC 3 | −0.04 (3.62) | −0.682 (3.38) | 2.65 (1.99) | 3.45 (1.76) | n.s. | n.s. |
| | | −1.18 (3.80) | | 2.01 (1.96) | | n.s. |
| MFCC 4 | −7.84 (3.76) | −6.16 (3.26) | −8.89 (3.23) | −7.78 (3.01) | n.s. | n.s. |
| | | −10.5 (2.94) | | −9.78 (3.19) | | n.s. |
| MFCC 5 | −11.9 (2.22) | −11.7 (1.96) | −13.6 (3.60) | −13.8 (3.99) | n.s. | n.s. |
| | | −12.2 (2.63) | | −13.5 (2.63) | | n.s. |
| MFCC 6 | −2.48 (3.05) | −1.59 (3.17) | −1.24 (3.25) | −0.145 (3.47) | n.s. | n.s. |
| | | −3.90 (2.29) | | −2.11 (2.84) | | n.s. |
| MFCC 7 | −3.64 (3.80) | −1.95 (3.37) | −3.41 (2.39) | −1.96 (2.23) | n.s. | n.s. |
| | | −6.29 (2.85) | | −4.58 (1.83) | | n.s. |
| MFCC 8 | −5.80 (3.27) | −5.84 (3.22) | −5.06 (2.69) | −4.91 (2.71) | n.s. | n.s. |
| | | −5.74 (3.46) | | −5.18 (2.74) | | n.s. |
| MFCC 9 | −2.91 (3.22) | −2.16 (3.27) | −2.62 (2.79) | −2.01 (3.05) | n.s. | n.s. |
| | | −4.09 (2.86) | | −3.11 (2.53) | | n.s. |
| MFCC 10 | −0.532 (2.67) | 0.658 (2.21) | −0.408 (2.44) | 1.28 (2.06) | n.s. | n.s. |
| | | −2.40 (2.28) | | −1.76 (1.83) | | n.s. |
| MFCC 11 | −2.33 (2.80) | −0.706 (2.05) | −2.57 (2.29) | −1.77 (2.46) | n.s. | n.s. |
| | | −4.89 (1.67) | | −3.21 (1.99) | | n.s. |
| MFCC 12 | −2.06 (1.97) | −1.96 (1.97) | −2.80 (1.80) | −3.01 (1.75) | n.s. | n.s. |
| | | −2.21 (2.02) | | −2.62 (1.87) | | n.s. |

SD: standard deviation.
n.s.: not significant after bonferroni correction.

with a telephone reciever type headset. We developed an in-house Android application which gives both audio and visual instruction of each tasks to participants and records their voices. Participants operated this application by themselves according to the guidance and read out or spoke as instructed. Voices were recorded using 16-bit PCM at 22.05 kHz.

### 2.3. Extracting acoustic features from the voice

Prior to processing, silent segments before and after utterances were manually removed from recordings. Voices were analyzed with OpenSMILE v2.1.0 (Eyben et al., 2013) using the 'Feature set of Interspeech 2009 Emotion Challenge' preset configuration (IS09_emotion.conf). This preset configuration was used to detect emotion in the voice at the Interspeech 2009 conference (Schuller et al., 2009). Using this configuration, the mean values of the following acoustic features within utterance segments were calculated: root mean square of energy, twelve dimensions of mel-frequency cepstram coefficient (MFCC), zero crossing rate, harmonics to noise ratio, and fundamental frequency.

### 2.4. Statistics

First, we compared each acoustic feature derived from voices between depression patients and controls using Student's $t$-test for all subjects, subjects divided by gender, and subjects divided by age (up to 40 years old and over 40 years old). Then, based on our hypothesis that some of the MFCCs could be descriminant factors between depressed patients and controls, MFCC features underwent stepwise discriminant analysis. Receiver Operatorating Characteristic (ROC) curve was also calculated to get its area under curve (AUC), sensitivity, and specificity. Following discriminant analysis, correlation analysis between acoustic features and QIDS-SRJ score was performed to explore the correlation between depressive symptoms and various acoustic features in depression patients or controls. Next, multiple regression analysis was performed to exclude affects by age and gender. Age, gender, group (depression patinets or controls) were set as explanatory variables, and acoustic features that were useful in discriminating between depression patients and controls were set as criterion variables. Finally, acoustic features that were useful in discriminating between depression patients and controls were compared before and after the VFT for the purpose of examining the reproducibility of changes in acoustic features.

All analyses were conducted with IBM SPSS Statistics 22 (IBM Corp, Armonk, NY, USA).

## 3. Results

### 3.1. Acoustic feature differences between groups

Student's $t$-test revealed that MFCC 2 was relatively higher in depression patinets than that in controls after Bonferroni correction for multiple comparisons ($t$ = −6.728, $p$ = 5.33 × $10^{-9}$, Cohen's $d$ = −1.72). No significant differences in other features were found. When subjects were divided by gender, MFCC 2 was also significantly higher in patients than in controls in each gender (male: $t$ = −4.903, $p$ < 0.001, Cohen's $d$ = −1.63; female: $t$ = −5.122, $p$ < 0.001, Cohen's $d$= −1.81) (Table 2). Subjects who were 40 years old or younger (younger group) consisted of 17 depression patients (male/female = 12/5) and 20 controls (male/female = 12/8), and subjects over 40 years old (older group) consisted of 19 depression patients (male/female = 10/9) and 16 controls (male/female = 4/12). In both age groups, MFCC 2 was significantly higher in depressive patients than in controls (younger group: $t$ = −5.327, $p$ < 0.001, Cohen's $d$ = −1.80; older group: $t$ = −4.858, $p$ < 0.001, Cohen's $d$ = −1.69). Shapiro-Wilk normality test of MFCC 2 gave $W$ = 0.97987, $p$-value = 0.3028, which implicated that MFCC 2 was similar to normal distribution.

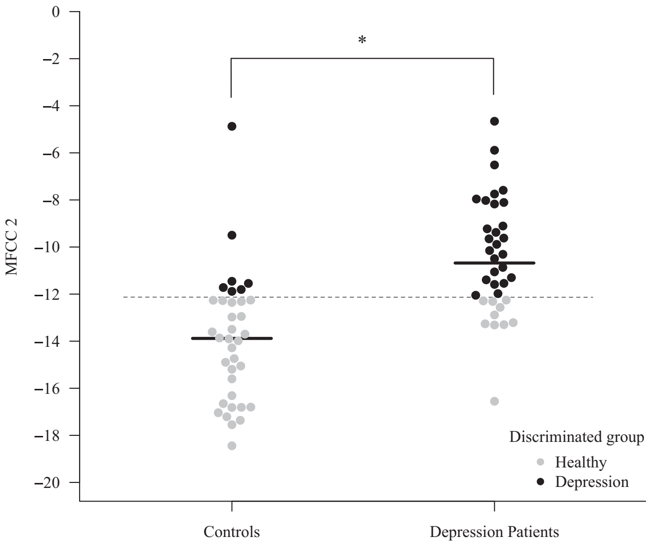

**Fig. 1.** MFCC 2 differences between Depression patients and Controls. Black dots indicates subjects discriminated as depression patients and grey dots indicate those discriminated as controls. Horizontal bar of each group shows average. * $p$ < 0.001.

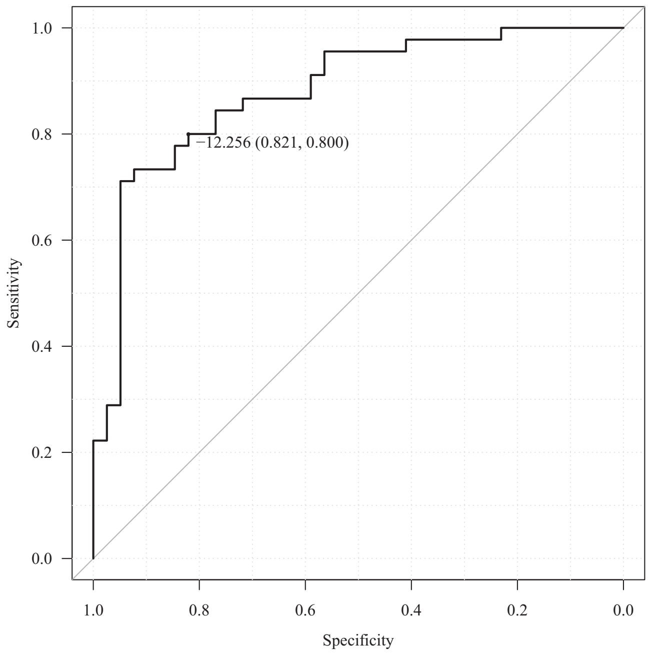

**Fig. 2.** Receiver Operatorating Characteristic (ROC) curve of MFCC 2. Area under curve (AUC) was 0.88.

**Table 3**
Discriminant analysis for MFCC 2.

| Groups | Results | |
|---|---|---|
| | Depression patients (male/female) | Controls (male/female) |
| Depression patients ($n$ = 36) | 28 (18/10) | 8 (4/4) |
| Controls ($n$ = 36) | 5 (3/2) | 31 (13/18) |

Stepwise discriminant analysis with MFCCs revealed that only MFCC 2 contributed to the discrimination (Wilks' $l$ = 0.567; $c^2$ = 39.456, $df$ = 1, p < 0.001).

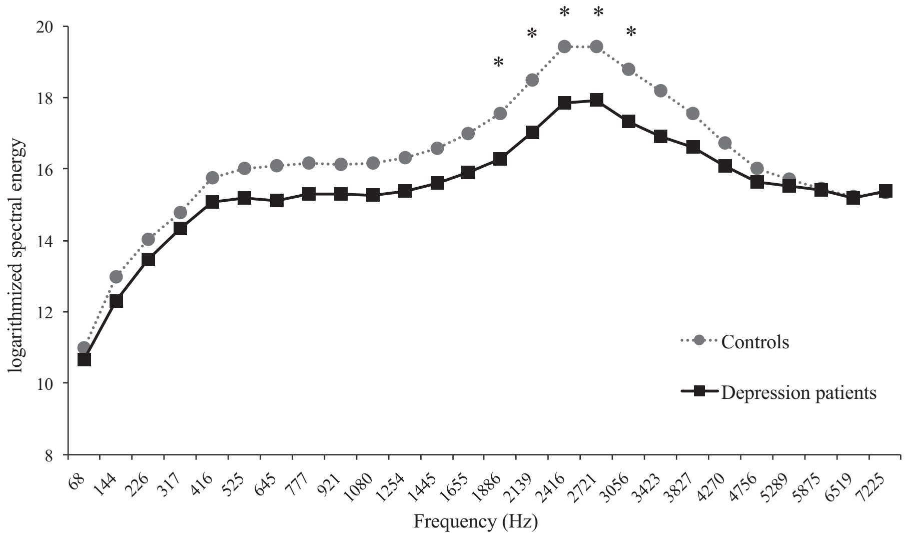

**Fig. 3.** Voice spectra of patients with depression and controls. Depression group voice had lower energy, around 2000–3000 Hz, than control group voices. * $p < 0.001$ corrected for multiple comparison with bonferroni correction.

### 3.2. Discriminant analysis with MFCCs

Stepwise discriminant analysis with MFCCs revealed that only the second dimension of MFCC (MFCC 2) significantly contributed to discrimination between depression patients and controls (Wilks' $\lambda = 0.567$; $\chi^2 = 39.456$, $df = 1$, $p < 0.001$). Using MFCC 2, sensitivity was 77.8%, specificity was 86.1%, and accuracy was 81.9% between depression patients and controls (Table 3). Fig. 1 shows beeswarm plots of MFCC 2 for each group. The unmedicated patient was correctly discriminated as a patient. Fig. 2 shows ROC curve of MFCC 2 and AUC was 0.88. Its sensitivity was 80.6% and its specificity was 83.3%.

### 3.3. Spectral difference between groups

MFCC 2 results can be explained by the broad shape of the voice spectrum (Mitrović et al., 2010). We compared the voice spectra of both groups. Fig. 3 shows the voice spectra that produced the MFCCs. The horizontal axis represents frequency scaled to the mel scale, and the vertical axis represents logarithmized spectral energy. The values represent the average of each group. In comparison with controls, the depression group had low energy, around 2000–3000 Hz, which was significantly different between groups after bonferroni correction. Fig. 4 shows the subtraction of the voice spectrum of patients with MDD from that of controls with basis functions of MFCC 0–3. The shape of the subtraction pattern was similar to that of the basis function of MFCC 2, which could explain the relationship between changes in MFCC 2 and the low energy in the depression group.

**Table 4**
Multiple regression analysis of MFCC 2.

| Variable | *r* | | | *ß* | *t* | *p-value* |
|---|---|---|---|---|---|---|
| | MFCC 2 | Gender | Age | | | |
| Group | 0.658 | −0.167 | 0.215 | 0.637 | 6.673 | $5.48\times10^{-9}$ |
| Age | 0.197 | 0.209 | | 0.068 | 0.708 | n.s. |
| Gender | −0.131 | | | −0.039 | −0.404 | n.s. |

n.s.: not significant.

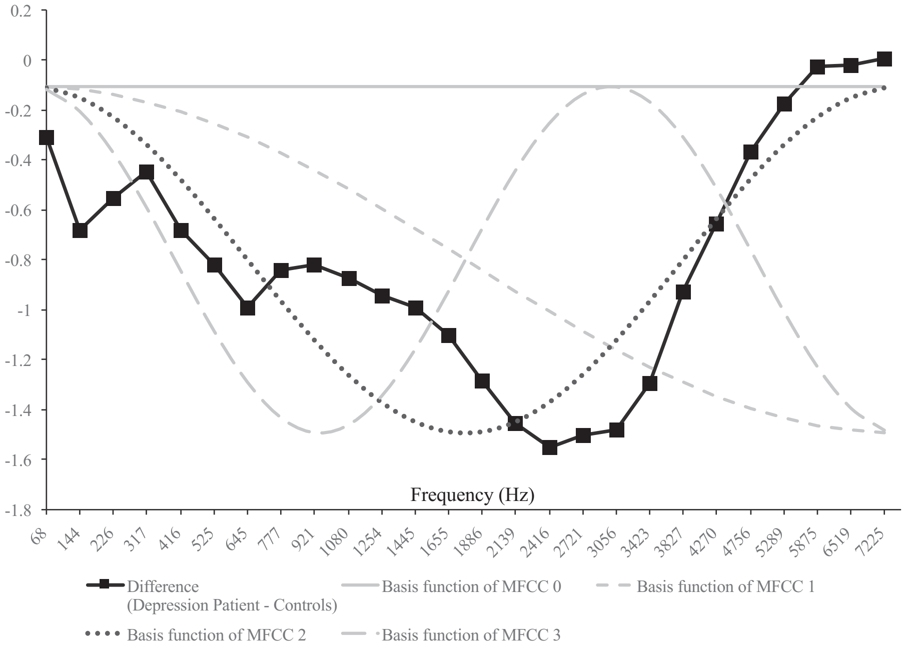

**Fig. 4.** MFCC 2 difference between depression patients and controls, with discrete cosine transform basis functions of MFCC 0–3. MFCC 2 difference resembled the basis function of MFCC 2. That is because the MFCC 2 difference was significant.

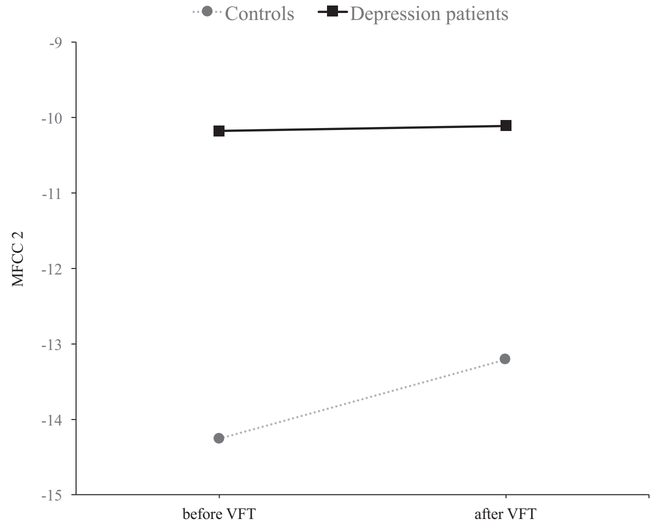

**Fig. 5.** MFCC 2 values before and after verbal fluency task. We did not find an interaction between temporal changes of MFCC 2 before and after the verbal fluency task.

### 3.4. Correlation analysis between acoustic features and QIDS-SRJ score

No acoustic features were correlated to QIDS-SRJ after Bonferroni correction. Correlation analysis between MFCC 2 and QIDS-SRJ resulted in $r = -0.023$, $p$-value $= 0.893$ in depression patients and $r = 0.25$, $p$-value $= 0.136$ in controls.

### 3.5. Multiple regression analysis of MFCC 2

Explanatory variables were set to age, gender, group (depression patinets or controls) and criterion variable were set to MFCC 2. Group's *ß* was 0.637, age's *ß* was 0.068, gender's *ß* was −0.039 and *R square* was 0.438 (Table 4).

### 3.6. ANOVA with repeated measures before and after VFT

MFCC 2 was significantly different in the group comparison, selected in discriminant analysis among MFCCs, and correlated significantly with QIDS-SRJ. Therefore, we used MFCC 2 for ANOVA with repeated measures before and after the VFT. Interaction between the VFT and group was not significant ($F = 3.511$). The VFT main effect within subjects was significant ($F = 4.508$, $p < 0.05$) and the main effect of group was also significant ($F = 51.117$, $p < 0.001$). Fig. 5 shows the average values of MFCC 2 for each group before and after the VFT.

## 4. Discussion

In this study, we found that MFCC 2 was relatively higher in depression patients than that in controls. This MFCC 2 difference was not affected by age or gender. These results suggest that changes in MFCC 2 might be caused by depression. However, MFCC 2 was not correlated to severity of depression.

### 4.1. Why is MFCC 2 different between groups?

Though there have been reports showing a relationship between MFCCs and depression, no reports have investigated the relationship between each MFC coefficient, especially MFCC 2, and depression. MFCCs have been shown to reflect vocal tract changes (Yinghua Zhu et al., 2013) and have been widely used in speech recognition. Since lower MFCC dimensions are not influenced by the vocal cords, these results could reflect vocal tract changes. The vocal tract, which consists of the pharynx, larynx, nasal cavity, and the oral cavity from vocal cords to lips, affects the timbre of the voice due to teeth, muscle tonus or movement of the tongue and lower jaw. Yinghua et al. showed that MFCC is related to the vocal tract using MRI (Yinghua Zhu et al., 2013).

In this study, we found that MFCC 2 corresponded to the lower voice spectral energy, around 2000–3000 Hz, in depression patients, which contributed to the discrimination of depression patients from control with more than 80% accuracy. Generally, this frequency band affects the projection of the human voice (Bele, 2006; Leino et al., 2011; Sundberg, 2001), and these results may reflect that the voice of the depression patient sounds muffled. In addition, frequency band of 2000–3000 Hz is most highly sensitive in normal equal-loudness-level contours (ISO 226:2003) which are perceived as equally loud by human listeners. As far as we know, no previous studies have shown the relationship between MFCC 2 and a lower voice energy spectrum of around 2000–3000 Hz in depression patients. Considering this, our results suggest that the frequency band of 2000–3000 Hz which is most highly sensitive in hearing could change in depression, leading to a muffled voice or changes in MFCC 2 in terms of acoustic features, and, thus, that MFCC 2 could be a biomarker for depression. For example, a voice sample from a telephone conversation may be used for objective psychiatric assessments using MFCC 2 without necessitating a clinical interview.

There remains the question whether MFCC 2 changes by treatments of depression. From our results, MFCC 2 did not correlate with severity of depression, which implies MFCC 2 may not reflect severity but depressive states itself. Future research is needed to clarify whether treatments of depression affect MFCCs.

### 4.2. Fundamental frequency

Although many studies have reported a relationship between fundamental frequency and psychological states including depression, our results showed no significant difference between depression patients and controls in this study. Fundamental frequency reflects pitch fraction and it has wide individual variation. We analyzed the average fundamental frequency across voice samples. Our voice samples, sets of “012–345–6789”, had silent intervals between each digit, which could have affected the values for the fundamental frequency feature; so, these values might be far from the actual fundamental frequencies. Further study is needed to determine whether fundamental frequency is useful for detecting depression.

### 4.3. Limitations

There are some limitations in this study. First, our sample size was small and more participants are needed in order to improve the reliability of these results. Second, though we used a telephone reciever type headsets to minimize environmental differences, we could not eliminate all background noises. However, MFCC itself is insulated from such noise. Moreover, we recorded the voices of the same subjects with various backgrounds and we did not observe significant differences in MFCC (data not shown). The effect of sound recording condition (with headset or not), sound recording program (sampling frequency or bit depth) is unclear. We are going to examine voices recorded in different sound recording condition or program in another study. Lastly, psychotropics could have a confounding effect on voice. We must study the voices of drug-naïve patients with depression in the future.

### 4.4. Further studies

We are going to examine usefulness of MFCC 2 for differential diagnosis of various depressive states including major depressive disorder or bipolar disorder. Moreover, for investigating changes of acoustic features by treatments, we are pushing forward the study to compare the voice before and after electroconvulsive therapy (ECT).

### 4.5. Conclusion

In this study, we showed that an acoustic feature of the voice, MFCC 2, changes in patients with depression, and that this change reflects the frequency band of 2000–3000 Hz. This change of MFCC 2 corresponds to clinical impressions that patients with depression have muffled voices. MFCC 2 could be useful in making a diagnosis of depression.

## Conflict of interest

Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masafumi Nishimura, and Tetsuaki Arai report no biomedical financial interests or potential conflicts of interest.

Masayuki Suzuki, Toru Nagano, and Ryuki Tachibana are employees of IBM Japan, LTD.

## Author disclosure

### Role of funding source

This study was supported by a Grant-in-Aid (24650417) from the Ministry of Education, Science, Sports and Culture of Japan.r

### Authors' contributions

Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto and Tetsuaki Arai designed the study and wrote the protocol. Takaya Taguchi, Hirokazu Tachikawa and Kiyotaka Nemoto managed the literature searches and collected voice samples. Masayuki Suzuki, Toru Nagano, Ryuki Tachibana and Masafumi Nishimura undertook voice data analysis. Takaya Taguchi, Hirokazu Tachikawa and Kiyotaka Nemoto undertook the statistical analysis, and Takaya Taguchi wrote the first draft of the manuscript. All authors contributed to and have approved the final manuscript.

## Acknowledgments

We are grateful to Dr. Tetsuya Takao (Mito Mental Clinic), Dr. Nagafumi Doi (Ibaraki Prefectural Medical Center of Psychiatry), and Dr. Shuzo Abe (Kurita Hospital) for their support.

## References

Akkaralaertsest, T., Yingthawornsuk, T., 2015. Comparative analysis of vocal characteristics in speakers with depression and high-risk suicide. Int. J. Comput. Theory Eng. 7, 448–452.

Alpert, M., Pouget, E.R., Silvia, R.R., 2001. Reflections of depression in acoustic measures of the patient's speech. J. Affect. Disord. 66, 59–69.

Bele, I.V., 2006. The speaker's formant. J. Voice 20, 555–578.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, P.J., 2004. Voice acoustical measurement of the severity of major depression. Brain Cogn. 56, 30–35.

Cummins, N., Epps, J., Breakspear, M., Goecke, R., 2011. An investigation of depressed speech detection: Features and normalization, In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. pp. 2997–3000.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F., 2015. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 71, 10–49.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Signal Process. 28, 357–366.

Eyben, F., Weninger, F., Groß, F., Schuller, B., Gross, F., Schuller, B., 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In: Proceedings of the 21st ACM International Conference Multimed. (MM 2013) pp. 835–838.

Fujisawa, D., 2010. Assessment scales of cognitive behavioral therapy. Jpn. J. Clin. Psychiatry 39, 839–850.

Herrmann, M.J., Ehlis, A.C., Fallgatter, A.J., 2003. Frontal activation during a verbal-fluency task as measured by near-infrared spectroscopy. Brain Res. Bull. 61, 51–56.

Inada, T., Inagaki, A., 2015. Psychotropic dose equivalence in Japan. Psychiatry Clin. Neurosci. 69, 440–447.

Joshi, J., Göcke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., Breakspear, M., 2013. Multimodal assistive technologies for depression diagnosis and monitoring. J. Multimodal User Interfaces 7, 217–228.

Ladd, D.R., 1980. Structure of Intonational Meaning: Evidence from English. Indiana University Press.

Leino, T., Laukkanen, A.M., Radolf, V., 2011. Formation of the actor's/speaker's formant: a study applying spectrum analysis and computer modeling. J. Voice 25, 150–158.

Low, L.-S.A., Maddage, N.C., Lech, M., Sheeber, L.B., Allen, N.B., 2011. Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans. Biomed. Eng. 58, 574–586.

Mitrović, D., Zeppelzauer, M., Breiteneder, C., 2010. Features for content-based audio retrieval. Adv. Comput. 78, 71–150.

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S., 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguist. 20, 50–64.

Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R., 2012. Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiatry 72, 580–587.

Nilsonne, A., Sundberg, J., Ternström, S., Askenfelt, A., 1988. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Am. 83, 716–728.

Pu, S., Nakagome, K., Yamada, T., Yokoyama, K., Matsumura, H., Mitani, H., Adachi, A., Nagata, I., Kaneko, K., 2012. The relationship between the prefrontal activation during a verbal fluency task and stress-coping style in major depressive disorder: a near-infrared spectroscopy study. J. Psychiatr. Res. 46, 1427–1434.

Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Depression 54, 573–583.

Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 emotion challenge. INTERSPEECH-2009, pp. 312–315.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Christian, M., Narayanan, S., 2010. The INTERSPEECH 2010 paralinguistic challenge. INTERSPEECH-2010, pp. 2794–2797.

Sundberg, J., 2001. Level and center frequency of the singer's formant. J. Voice 15, 176–186.

Tolkmitt, F.J., Scherer, K.R., 1986. Effect of experimentally induced stress on vocal parameters. J. Exp. Psychol. Hum. Percept. Perform. 12, 302–313.

Wittels, P., Johannes, B., Enne, R., Kirsch, K., Gunga, H.-C., 2002. Voice monitoring to measure emotional load during short-term stress. Eur. J. Appl. Physiol. 87, 278–282.

Yinghua Zhu, Yoon-Chul Kim, Proctor, M.I., Narayanan, S.S., Nayak, K.S., 2013. Dynamic 3-D visualization of vocal tract shaping during speech. IEEE Trans. Med. Imaging 32, 838–848.