2019

Classifying females' stressed and neutral voices using acoustic-phonetic analysis of vowels: an exploratory investigation with emergency calls

Tavi, Laura

Springer Nature America, Inc

Tieteelliset aikakauslehtiartikkelit

http://dx.doi.org/10.1007/s10772-018-09574-6

International Journal of Speech Technology
https://doi.org/10.1007/s10772-018-09574-6

Classifying females' stressed and neutral voices using acoustic–phonetic analysis of vowels: an exploratory investigation with emergency calls

Lauri Tavi[1]

Received: 5 February 2018 / Accepted: 15 November 2018

Abstract
In the present exploratory study, we investigated acoustic–phonetic measures of spoken vowels for detection of female speech under conditions of stress. Eight authentic recorded calls to emergency services received from eight Finnish adult female speakers were chosen for the analysis. Based on the purpose of the call, the recordings were divided into two groups: the stressed group and the neutral group. We chose f0, H1–H2 and centre of gravity as acoustic–phonetic predictors for our final classification models; In previous studies, high f0 has been associated with a stressed voice, but H1–H2 and centre of gravity have not previously been related to speech under stress. On the other hand, H1–H2 has been used to detect non-modal voice qualities, such as a creaky or breathy voice, and similar voice qualities have been observed in stressed speech. Furthermore, indications exist that in speech under stress, acoustic energy is concentrated in higher frequencies, which consequently increases the centre of gravity. We tested stress detection accuracy with three statistical classifiers: LDA, logistic regression and decision tree. Our results indicated that all the models performed better when they were trained using only the vowel /i/ rather than training them with all Finnish vowels. The use of our best performing model, a logistic regression model based on /i/, yielded 88% correct classification, whereas a logistic regression model trained with all vowels achieved an accuracy of only 81%. We conclude that the results indicate a good stress classification accuracy, although further research with more extensive data is required.

Keywords Speech under emotional stress · Acoustic–phonetic analysis · Emergency calls · Finnish vowels · Female voice

1 Introduction

Speaker classification systems and emotion-related speech feature extraction models have various, yet closely related aims: identification and verification of individuals for forensic purposes, ensuring security-protected access, improving speech recognition and synthesis systems, and also developing methods to reveal the physical or psychological state of the speaker (Hill 2007). For example, some automatic speaker verification systems have already outperformed human listeners (Hautamäki et al. 2010), although human listeners are more sensitive to subtle emotional changes in speech than human–machine interfaces (Hansen and Patil 2007). Nonetheless, different types of supporting automatic speaker identification and profiling systems already exist; these systems commonly use mel-frequency cepstral coefficients (MFCCs) for the extraction of speaker-dependent features, and for modelling they utilise e.g., linear discriminant analysis, support vector machines, gaussian mixture, hidden Markovs and neural nets models (Gałka et al. 2015; Ververidis and Kotropoulos 2006; Steeneken and Hansen 1999).

From the perspective of speech science, one significant approach to examine speaker specific features has been the study of speech prosody. Prosodic features of speech consist of variations in intonation (f0), speech or articulation rate, and loudness or intensity. Along with energy, f0 is the most widely used acoustic feature, especially in the analysis of emotions (Cummings et al. 2015). However, in recent studies, speech rate and rhythm have also been found to be promising features for speaker identification (Cummings et al. 2015; Dellwo et al. 2015).

As Farrús (2008) pointed out, recognition systems based on only prosodic features do not generally outperform traditional

✉ Lauri Tavi
lauri.tavi@uef.fi

[1] School of Humanities, University of Eastern Finland, Yliopistokatu 4, Agora, Joensuu, Finland

Published online: 27 November 2018

filter-based systems, although they have been successfully used to improve the performance of the traditional systems. However, more study is required to understand acoustic and prosodic correlates of emotions in speech. Especially from the phonetic and linguistic point of view, the inclusion of natural speaker context, such as a speaker's dialect or quality of conversation, in speech analysis is important in order to gain new information about speech production and acoustics.

The aim of this exploratory study is to distinguish female voice under stress conditions from neutral female voice without the emergency-related factor based on acoustic analysis of specific vowels. In emotion classification models, focusing on linguistically annotated data is necessary since, as Meyer et al. (2018) have reported, even the state-of-the-art deep learning classifiers might actually learn only linguistic content instead of desired emotions. Thus, we investigated acoustic–phonetic features of speech under psychological stress using 1792 vowels from eight authentic emergency call recordings and tested the classification accuracy with three different statistical methods. This paper is structured as follows: The following section introduces the concept of speech under stress. Speech recordings, measurement techniques and statistical analyses are described in detail in the Sect. 3. The findings are discussed in Sects. 4 and 5.

2 Speech under stress

During recent decades, voice stress extraction has been a popular research subject in the field of speech analysis. The interest in voice stress extraction is due to e.g., developments in the methods of forensic phonetics and security access applications (Jessen 2008). Furthermore, stressed speech has been noted to have a negative effect on the accuracy of speech recognition systems (Hansen and Patil 2007). Previous studies have shown, however, that the term psychological stress is problematic to define; this is one reason why different studies have analysed speech under stress from various data sets and reported inconsistent results concerning acoustic correlates of stress (Kirchhübel et al. 2011).

He et al. (2011) classified three different types of data sets that have been used in previous studies of speech under stress: (1) emotions simulated by professional actors, (2) experimentally induced emotional expressions in a recording laboratory and (3) natural vocal expressions, such as emergency calls, recorded in the field. Therefore, acoustically measured stress can be a result of an actor's interpretation, a time-measured cognitive task or natural fear of death. All of these data sets have advantages and disadvantages; for instance, Demenko (2008) has pointed out that

> studies using actors and simulated stress or emotions have the advantage of a controlled environment. The major disadvantage is, however, an artificial experimental design which can result in producing highly exaggerated misrepresentations of emotions in speech. Another group focuses on the analysis of authentic recordings coming from actual situations. There is usually no doubt as to the presence of stress in these recordings, however there is a problem of categorization of the homogeneous classes of stress.

Along with the challenges of different qualities of stress, Hollien (1990) described another methodological issue concerning analysis of speech under stress: the level of the stress. The volume of emotions varies during a state of mind categorizable as "stressed", rather than being fully absent or constant. However, quantitative measures of the level of stressed speech, or at least comparisons of these measures between different conditions, might be overly complex to take into consideration. Due to the difficulty of quantification of emotions, in the studies of emotional speech research material requires detailed selection and description, which might be a prominent reason why emotion recognition databases have sometimes only ten or less speakers (Jacob 2017; Milošević and Đurović 2015).

Despite the somewhat ambiguous nature of the concept of speech under stress, it is still useful for detecting "stressed" from "non-stressed" speech; for example, separating calls to emergency services that report a real emergency situation (i.e., stressed speech) from those that do not report any emergency (i.e., non-stressed speech) might be a decision of life and death for a duty officer. In the present study, we classify speech from authentic emergency call recordings that included a direct health or life hazard as emergency related speech under stress (ERSUS).

2.1 Stress and articulatory system

Speech production requires complex articulatory movements and controlled airflow from the respiratory system, which are also sensitive to certain emotional situations (Hansen and Patil 2007). Regarding psychological stress, previous studies have reported an increase in respiration rate. The increased breathing raises the sub-glottal pressure, which leads to glottal pressure through the supra-laryngeal vocal tract and causes friction and turbulence in the voice (Kirchhübel et al. 2011). Especially female voices under stress have been observed to be more breathy and strained (Van Lierde et al. 2009).

In addition to the increased respiration rate, previous studies have reported muscular tensions in the vocal tract during stress (Hansen and Patil 2007; Steeneken and Hansen 1999), which causes further variability of airflow characteristics (Zhou et al. 2001). Muscular tension can also have a restrictive effect on the articulatory system, and one possible consequence of this is vowel centralization (Tavi 2017).

Furthermore, voicing irregularities or "voice breaks" have been observed to occur in stressed speech (Kirchhübel et al. 2011), which might also be caused by the increase in respiration rate or muscular tension. However, to the best of the author's knowledge, little empirical information exists about the effects of emergency-related stress on the articulatory system.

2.2 Acoustic correlates of stress

Numerous acoustic parameters have been considered as possible sensors of speech under stress, such as fundamental frequency, jitter, shimmer, intensity, duration, and formants (Tavi 2017; Sondhi et al. 2015). Comprehensive overviews of acoustic correlates to vocal stress can be found in Kirchhübel et al. (2011) or Jessen (2006). This section focuses on three measures that we have used for stress detection purposes in our investigations. In addition to the commonly used f0, we present two other previously uncommon stress correlates: the difference between the amplitude peaks of the first and second harmonics (H1–H2) and centre of gravity.

Previous studies have emphasized the importance of f0 values in the perception of psychological stress (Sondhi et al. 2015; Demenko and Jastrzebska 2012; He et al. 2011). In these studies, high f0 values are strongly related to stress voice, although contrary results also exist (Van Lierde et al. 2009). In addition, Kirchhübel et al. (2011) stated that mean f0 appears to increase more in real-life stress than in experimentally induced stress in a laboratory.

H1–H2 is one of the spectral tilt measures. Although the difference between the amplitudes of the first and the second harmonics is not a common stress correlate, it has been used to measure different non-modal voice qualities, for example breathiness (Keating and Esposito 2006), or creaky and tense voice (Kreiman and Gerratt 2010). Moreover, similar voice qualities have been observed in speech under stress (Van Lierde et al. 2009). However, since Simpson (2012) reported gender-related differences in H1–H2 when measuring breathiness, potential stress-related variation in H1–H2 could also be gender-specific.

In comparison to neutral voice, indications exist that in speech under stress, acoustic energy is concentrated in higher frequency bands (Kirchhübel and Howard 2013). In phonetics, centre of gravity (CoG) is considered as the mean frequency of the spectrum; the more the energy of the speech signal is concentrated to higher frequencies, the higher is CoG. This provides the theoretical motivation for measuring CoG for stress detection purposes, even though CoG has not previously been included in vocal correlates to stress.

Overall, previous studies have used assorted measurement techniques for acoustic stress measures, which along with the different data sets is another reason why the earlier findings with regard to the acoustic correlates of stress voice are rather inconsistent (Kirchhübel et al. 2011; Harnsberger et al. 2009). For example, Streeter et al. (1983) reported that no reliable and valid acoustic indicators of psychological stress exist, and Kirchhübel and Howard (2013) suggested that instead of referring to the reliable acoustic indicators of emotions, "it is more appropriate to regard them as acoustic tendencies". Since the results from different stress voice studies have been rather inconsistent, the reported acoustic correlates of stress might be restricted to specific data sets; therefore, for example, the acoustic measurements from authentic stress voice might be incompatible with the acoustic measurements from simulated stress.

3 Materials and methods

3.1 Speech data

In this study, the research material consisted of eight authentic Finnish emergency call recordings from eight female speakers. All the emergency calls were received one day in the year 2016. The recordings were collected from the Kuopio Emergency Response Centre in Northern Savonia, Eastern Finland, because of the expectation that calls received from Eastern Finland would have less linguistic diversity than calls from other parts of Finland, where e.g., Swedish is also spoken.

We categorised four callers' recordings as ERSUS (emergency related speech under stress) since they all included a citizen's report of a direct life or health hazard and we observed them sounding emotionally stressed during the whole call.[1] The other four recordings were categorized as neutral speech; in this category the callers were officials (i.e., duty officer, nurse, home aid and worker from child welfare), and the purpose of the call was a work assignment, without any direct life or health hazard. Another reason for selecting these particular official callers was that we observed no audible emotional stress from these callers. In addition, both categories were restricted to young adult females without any conspicuous speech features.

Table 1 above shows a detailed description of the speech data. Although the speakers had some breathiness and creakiness in their voice, we selected only speech segments that we considered as rather modal type; approximately 10–20% of the speech was excluded from the analysis. Thus, creaky voice, whisper, or screaming do not occur in the analysed data.

[1] Even though a direct life or health hazard causes stress, it is possible that a caller's stress level changes during a call; however, in this data set we considered it to be implausible that during the phone call (which lasted approximately less than 2 min) emotion of stress caused by a serious emergency situation could vanish entirely. Thus, we decided to include all clear sounding vowels in the analysis.

Table 1 Description of the speech data

Caller	Group	Purpose for the emergency call
Young adult female A	Stress	A caller reports that she nearly hit a confused man with her car and is now concerned of his safety
Young adult female B	Stress	A caller reports that she saw a man strangling another man in a street
Young adult female C	Stress	A caller reports her ex-boyfriend's suicide attempt
Young adult female D	Stress	A caller reports that a man is threatening her
Young adult female E	Neutral	A nurse requests a police to come to get the DNA sample and to question a crime victim
Young adult female F	Neutral	A nurse calls an ambulance without urgency for elderly person
Young adult female G	Neutral	A child welfare official reports a missing adolescent
Young adult female H	Neutral	A police calls an ambulance for a person who might need mental care services

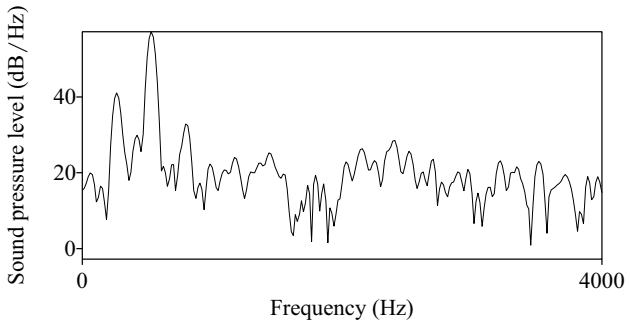

Fig. 1 Spectral slice of /i/-vowel from speaker D. Frequency of the first harmonic (i.e., f0) is 266 Hz (the second highest peak) and of the second harmonic is 532 Hz (the highest peak)

Table 2 Speakers' vowel counts

Caller	Group	Number of /i/-vowels	Number of all vowels
A	Stress	64	370
B	Stress	73	374
C	Stress	26	224
D	Stress	29	200
E	Neutral	23	125
F	Neutral	18	158
G	Neutral	38	234
H	Neutral	17	107

3.2 Acoustic measurement technique

Telephone recordings are often an important data source in an accident or forensic investigation. However, telephone speech is challenging data for acoustic analysis if, for example, the quality of the telephone connection is weak, the telephone channel contains distortion or the speaker's distance from the microphone is too great. In addition, common sampling rate for telephone recordings is 8000 Hz since only frequencies less than 4000 Hz are transmitted by telephones (see Fig. 1).

The rejection of frequencies over 4000 Hz affects especially high frequency sounds such as sibilants (Niemi-Laitinen 1999); For example, previous studies have reported that the way of production of /s/ is related to the speaker's gender (Li et al. 2016), and even to male speakers' sexual orientation (Tracy et al. 2015). Therefore, necessary information for speaker profiling might be lacking from the call recordings due to limited transmitted frequencies.

We analysed authentic telephone recordings from real life emergency situations. Since the signal quality varies during the recordings, acoustic analyses were carried out only for clear sounding vowels. All vowels were annotated manually with Praat (Boersma and Weenink 2017) under the following conditions: (1) non-overlapping speech, (2) duration over 30 ms, (3) normal voice quality and (4) no loud background noise.

The Finnish language has eight vowels, which can occur short; /i/, /y/, /u/, /e/, /ø/, /o/, /æ/ and /a/, or long; /i:/, /y:/, /u:/, /e:/, /ø:/, /o:/, /æ:/ and /a:/. Finnish has also several different diphthongs and vowel sequences. However, we focused on short and long /i/-vowels, since previous studies have indicated that vowels can carry emotional information (e.g., Waaramaa et al. 2014) and focusing on a specific phoneme category should reduce inter and intra-speaker variation in acoustic measures and reveal potential acoustic differences between speech under stress and neutral speech (Tavi 2017). In addition, Niemi-Laitinen (1999) reported that based on the Euclidean distance, Finnish /a/ and /e/ have the highest interspeaker variation and /u/ and /i/ have the lowest variation.

Table 2 shows the speakers' vowel counts. We measured various acoustic parameters with Praat using two Praat scripts: ProsodyPro 5.6.3 (Xu 2013) and phonation-measurements (Vicenik 2017). Based on the preliminary measurements with the aforementioned Praat scripts, the following predictors were selected for statistical modelling: median f0 in Hz, H1–H2 in dB and CoG in Hz (see speakers' mean values in Table 3).

We chose f0 median value since we expected the median value to be a more reliable measure for the telephone quality

Table 3 Mean values of acoustic–phonetic measures by the speakers A–H

Measure	A (S)	B (S)	C (S)	D (S)	E (N)	F (N)	G (N)	H (N)	Group (S)	Group (N)
F0 (Hz)	215	248	250	262	218	216	187	223	239	206
H1–H2 (dB)	− 17.9	− 13.9	− 8.5	− 15.4	− 1.9	− 5.3	− 5.7	− 4.2	− 14.7	− 4.4
CoG (Hz)	928	550	589	731	511	531	432	469	709	476

The column Group (S) shows the mean values of the stress group and the column Group (N) shows the mean values of the neutral group

data, in comparison to e.g. maximum or mean f0. In addition, the developer of ProsodyPro, Yi Xu, associated shifts in median pitch with emotions such as fear (Xu et al. 2013) and used median pitch as a default in bio-informational dimension measurements in ProsodyPro.

Median f0 is calculated with ProsodyPro without any manual pulse corrections. Along with f0, CoG is measured with the same Praat script. Although ProsodyPro also calculates H1–H2, we measured H1–H2 with Vicenik's phonation-measurements script. The reason for the use of Vicenik's script is that Vicenik's H1–H2 measurements corresponded with our manual checking, whereas they conflicted with ProsodyPro's H1–H2 calculations. In addition, it should be noted that measurements of H1–H2 and CoG might not be comparable with the measurements of other speech recordings from different database since speech coding, or different codecs, affects to speech spectrum. Nevertheless, the measurements are comparable inside of the same data set; in the current study, we verified the correctness of acoustic measurements manually.

3.3 Statistical analyses

Three different classifiers were used for stress detection: linear discriminant analysis (LDA), logistic regression (LR) and decision tree (DT). Since this study is limited to between-speaker design for the reason that experimental within-speaker stress measurements were infeasible with previously recorded emergency calls, we trained and tested classifiers using the leave-one-out cross validation method, i.e., using all speakers one by one as a test data while the rest of the speakers served as a training data. All statistical analyses were carried out in R (R Core Team 2017).

LDA, LR and DT are commonly used methods to predict binary or polytomous categorical class using one or more predictors. In the following, the classifiers are described briefly; more in-depth coverage about using these classifiers can be found e.g., in (Piegorsch 2015).

LDA is a parametric technique for determining weightings of predictors in order to discriminate between two or more groups, and it is closely related to the analysis of variance. The aim of LDA is to find the best linear combination of features which separates the classes; however, LDA makes an assumption that dependent variables are normally distributed. (Piegorsch 2015.) In the present study, we used the LDA implemented in the MASS package (Venables and Ripley 2002).

Logistic regression is a standard regression model, which has been applied in numerous speech perception studies. The basic goal of LR is to fit a sigmoidal curve to categorical response data (Morrison and Kondaurova 2009; Morrison 2007). In comparison to LDA, logistic regression has the same advantages without the assumption of normal distribution (Morrison and Kondaurova 2009). We build the LR models using H2O package (H2O.ai team 2017).

A somewhat more advanced nonparametric method, the decision tree, is a model in the form of a tree structure which is built using recursive partitioning based on supervised classification rules (Piegorsch 2015). Decision trees can handle both categorical and numerical data; the model separates the data into smaller classes with decision nodes that are split into logical choices, and the result of the model is shown in terminal nodes or leaf nodes. (Lantz 2013.) For the decision trees, we used the Rpart (Therneau et al. 2017) and Rpart.plot (Milborrow 2017) packages.

All three classifiers were constructed using three predictors, f0, H1–H2 and CoG (see Acoustic Measurement Technique), in determining a binary speaker group, i.e., the stress group and the neutral group. Whereas distributions of H1–H2 are rather gaussian, f0 and CoG have skewed distributions. As a result, in the LDA models, logarithmic transformations were made for f0 and for CoG. Furthermore, since real emergency calls are longer than administrative calls, the stress and the neutral group have unequal numbers of vowels, i.e., of all the data 1/3 is from neutral callers and 2/3 is from stressed callers. Thus, we used balanced (i.e., 1/2 and 1/2) prior probabilities in all aforementioned models. After calculating the stress prediction accuracy for each classifier, we compared the correct prediction rates with the tests of proportions using Bonferroni adjusted alpha level.

4 Results

In this study, we investigated if focusing on a specific phoneme category yields better stress classification accuracy in comparison to analysing heterogeneous phoneme categories. Additionally, based on acoustic measurements of /i/-vowels

Table 4 Prediction accuracy of the LDA_i

	Stress /i/-vowels	Neutral /i/-vowels	Total
Correctly predicted	162	89	251
Falsely predicted	30	7	37
Accuracy%	84	93	87

The second and the third row show the number of predicted /i/-vowels and the final row shows the accuracies of correctly predicted /i/-vowels

Table 5 Prediction accuracy of the LDA_{all}

	All stress vowels	All neutral vowels	Total
Correctly predicted	893	526	1419
Falsely predicted	275	98	373
Accuracy%	76	84	79

The second and the third row show the number of predicted vowels and the final row shows the accuracies of correctly predicted vowels

Table 6 Prediction accuracy of the LR_i

	Stress /i/-vowels	Neutral /i/-vowels	Total
Correctly predicted	168	85	253
Falsely predicted	24	11	35
Accuracy%	85	89	88

The second and the third row show the number of predicted /i/-vowels and the final row shows the accuracies of correctly predicted /i/-vowels

Table 7 Prediction accuracy of the LR_{all}

	All stress vowels	All neutral vowels	Total
Correctly predicted	926	519	1445
Falsely predicted	242	105	347
Accuracy%	79	83	81

The second and the third row show the number of predicted vowels and the final row shows the accuracies of correctly predicted vowels

and of all vowels, we compared three different classification techniques; We trained and tested the LDA_i, the LR_i and the DT_i models with a total of 288 observations of Finnish short and long i-vowels from eight speakers. In the reference models, i.e., the LDA_{vowels}, the LR_{vowels} and the DT_{vowels}, we used a total of 1792 observations of all eight Finnish short and long vowels from the same speakers (see Sect. 3.2). For each model, we calculated the prediction accuracy for summed vowels from all speakers using the leave-one-out cross validation method. In addition, we used a $>50\%$ threshold value of correctly classified vowels within each speaker for deciding whether the speaker was predicted into the correct class.

The following sections present the stress classification accuracy for each classifier. The results from the classifiers are compared in Sect. 4.4.

4.1 Linear discriminant analysis

The LDA models were characterized with three acoustic variables: median f0, H1–H2 and CoG. Since median f0 and CoG have skewed distributions, we used logarithmic transformations for these variables. Tables 4 and 5 present the results from the LDA_i and the LDA_{vowels}, respectively. Table 4 shows that the observations of 288 /i/-vowels were predicted correctly with good accuracy especially for neutral /i/-vowels; 93% of /i/s were classified correctly. For stress vowels the prediction rate was also relatively high but less accurate, with 84% correct classification. The overall accuracy of the LDA_i was 87%.

In comparison to the LDA_i, the classification accuracy of the LDA_{vowels} was somewhat weaker; stress vowels were predicted correctly with an accuracy of 76% and neutral vowels with an accuracy of 84%. Hence, the overall accuracy of the LDA_{vowels} was 79%.

Tables 4 and 5 show that the correct prediction rate was lower when all vowels instead of a specific vowel category were under investigation. The difference between the overall accuracy of the LDA_i and of the LDA_{vowels} was also statistically significant at the five percent level ($p=0.002105$). In addition, by using a $>50\%$ threshold value of correctly classified vowels for deciding whether the speaker was under stress, the LDA_{vowels} classified one stress speaker falsely as neutral speaker; the correct prediction rate for speaker C was only 44% (see Table 1 for speaker's details). In the LDA_i model, these results from the same speaker was 62%. Thus, all speakers were categorized into the correct class only with the LDA_i model.

4.2 Logistic regression

In the LR models, we used the same predictors as in the LDA models: median f0, H1–H2 and CoG. As mentioned earlier, logistic regression makes no assumptions of normality of distribution, and hence there was no need for logarithmic transformation of f0 or of CoG.

The results from the LR_i and the LR_{vowels} are presented in Tables 6 and 7. Table 6 shows that the LR_i performed with a good accuracy; 85% of stress /i/-vowels and 89% of neutral /i/s were recognized correctly, with an overall accuracy of 88%.

However, when the LR model was trained based on acoustic measurements from all vowels, the classification

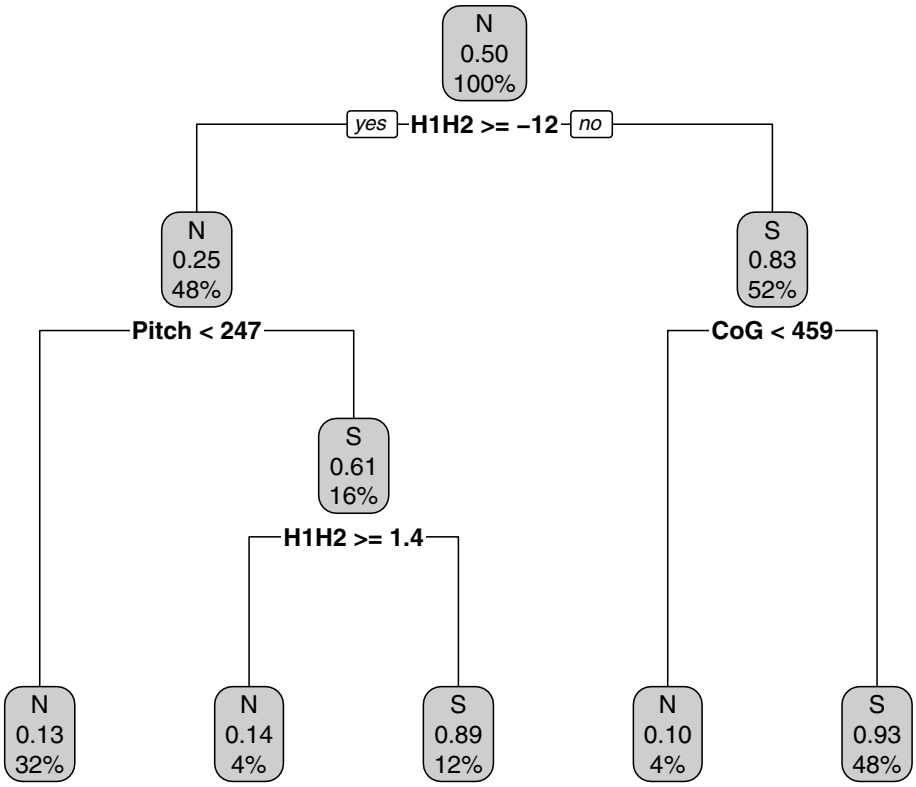

Fig. 2 Decision tree model. S stands for /i/ under stress and N stands for neutral /i/. f0 (*Pitch*), H1–H2 (*H1H2*), and centre of gravity (*CoG*) are used as dependent variables for class prediction

accuracy of stress vowels decreased from 85 to 79% and the correct prediction rate for neutral vowels decreased from 89 to 83%. The overall classification accuracy of the LR_{vowels} was 81% (see Table 7).

The results indicate that the LR model was more accurate for stress detection purposes when acoustic measurements were focused on /i/-vowels only; the difference between the proportions of the correct predictions of the LR models was statistically significant (p-value = 0.00435). In addition, the LR_{vowels} classified one stress speaker (C) falsely into the neutral category, whereas the LR_i detected the correct class for every speaker with a 50% threshold of correctly classified vowels.

4.3 Decision tree

The DT models were also characterized with median f0, H1–H2 and CoG. In comparison to LDA and LR, decision trees have an advance of visual representation, which is easy to plot and to understand (Piegorsch 2015). Figure 2 shows how the DT_i splits the data into decision nodes.

Stress detection accuracies of the decision tree models are presented in Tables 8 and 9. Table 8 shows that the DT_i classified stress and neutral /i/-vowels correctly with an accuracy of 76% and 73%, respectively. The overall correct prediction rate of the DT_i was 75%, which is the lowest percentage of all three /i/-vowel based classifiers.

Table 8 Prediction accuracy of the DT_i

	Stress /i/-vowels	Neutral /i/-vowels	Total
Correctly predicted	145	70	215
Falsely predicted	47	26	73
Accuracy%	76	73	75

The second and the third row show the number of predicted /i/-vowels and the final row shows the accuracies of correctly predicted /i/-vowels

Table 9 Prediction accuracy of the DT_{all}

	All stress vowels	All neutral vowels	Total
Correctly predicted	767	498	1265
Falsely predicted	401	126	527
Accuracy%	66	80	71

The second and the third row show the number of predicted vowels and the final row shows the accuracies of correctly predicted vowels

As Table 9 shows, DT_{vowels} performed slightly less strongly than DT_i, with accuracy of 66% for stress vowels, but 80% for neutral vowels, which is better accuracy than with DT_i. Yet, the overall accuracy of DT_{vowels} decreased to 71%, although the difference between the proportions of

Table 10 Classification matrix for the overall accuracy of three classifiers based on /i/-vowels

Classifier	Overall accuracy of all /i/-vowels (%)	Accuracy range between speakers (max%–min%)
LDA_i	87	100–62
LR_i	88	100–65
DT_i	75	94–58

The column on the right shows the classification accuracy range from maximum to minimum between speakers; the column shows that in each model, over 50% of every speakers' /i/-vowels are correctly classified

overall correct classifications of the DTs was not statistically significant ($p = 0.1796$). However, when using a $> 50\%$ threshold value for correctly classified speakers, DT_{vowels} predicted speaker C into the false class (see similar results in the Sects. 4.1 and 4.2; for this speaker, the accuracy of correctly predicted vowels was only 37%. On the other hand, the DT_i classified all speakers into the correct class, i.e., the prediction accuracy for correctly classified vowels were $> 50\%$ for all eight speakers. Thus, as in the LDA and the LR models, constructing the DT model on the basis of a specific vowel category was more efficient for stressed speaker detection than constructing the model on the basis of heterogeneous vowel categories.

4.4 Summary of the three classifiers

All three classifiers, i.e., linear discriminant analysis, logistic regression and decision tree, revealed a higher recognition rate when the models were based on acoustic analysis of /i/-vowels rather than of all vowels. Setting the threshold value of correctly predicted vowels within each speaker to $> 50\%$, /i/-based models classified all speakers in data set into the correct class, whereas each model based on heterogeneous vowels made one misclassification out of eight speakers. In addition to the fact that centre of gravity covaries with vowel quality, one explanation for this might be that focusing on a specific phoneme category simply reduces within-speaker variation in acoustic measurements and, consequently, reveals more effectively the acoustic differences in speech between emotional states. Of all the classifiers, the LR_i, which is formed as follows:

The LR_i model

$$logit\left(\frac{p}{(1-p)}\right) = -22.892 + 0.066f0 + 0.008CoG - 0.413(H1 - H2), \quad (1)$$

performed best, with an overall accuracy of 88 percent and with the highest maximum and minimum accuracy rate between speakers (see Table 10).

Tests of proportions show that the overall accuracy rate of the DT_i differed statistically significantly from that of the LR_i ($p < 0.001$) and that of the LDA_i ($p < 0.001$), whereas the differences in overall accuracy between the LR_i and the LDA_i ($p = 1$) was not statistically significant. Furthermore, Table 10 shows that the LR_i differed from the other models in its lower accuracy range between speakers; the LR_i had the highest maximum (100%) and the highest minimum (65%) speaker-specific /i/-vowel classification accuracy. Yet, all /i/-based models classified over 50% of /i/s into the correct class, which enabled the correct binomial stress/neutral classification for each speaker in the data set.

5 Discussion

Although this study was conducted with a relatively small database, the results support the following conclusions for ERSUS of young adult females:

1. Along with f0, CoG and H1–H2 can also be utilised to detect emotional stress in the speaker's voice
2. Instead of analysing heterogeneous phoneme categories, focusing on a specific phoneme category yields better stress classification accuracy
3. Traditional statistical models with a low computational cost seem to be an efficient stress voice classifier for a limited amount of speech data with a binary dependent variable

Another limitation in this study is the between-speaker design. However, since experimental stress measurements are infeasible with emergency call recordings, we used the leave-one-out cross validation technique and limited the data selection to young adult females without any conspicuous speaker characteristics, in order to minimise the natural inter-speaker variation.

As some previous studies have pointed out, building a robust or universal stress detection model based only on acoustic measures might be impossible to achieve (Kirchhübel et al. 2011; Streeter et al. 1983). Nevertheless, although no robust acoustic measure of stress exists, new exploratory combinations of acoustic parameters can still provide reasonably effective stress detection for a specific purpose (e.g., automatic pre-classification of emergency calls and applications in security access in the future), as well as supplementary information for human evaluation. In addition, further acoustic–phonetic analysis of stress voice will lead to a better insight into speech production in stressful situations.

6 Conclusion

In this study, we measured f0, CoG and H1–H2 from manually segmented vowels for classification of female speech under psychological stress in a special context; We investigated 1792 vowels from authentic call recordings to the emergency services from young adult females and tested stress classification accuracy using the leave-one-out cross validation method with three different statistical methods: LDA, logistic regression, and decision tree. The results showed that all models performed better when they were trained with acoustic measures from /i/-vowels rather than from heterogeneous vowel categories. Of all the classifiers, the logistic regression and the LDA model based on the /i/-vowel performed with the highest accuracy. We conclude that f0, CoG and H1–H2 appear to be a promising combination of acoustic measures for female stress voice detection from authentic emergency call recordings; However, since large numbers of authentic emergency call recordings are not usually available, for ethic or for legal reasons, we emphasize the fact that more investigation from larger data sets, where male speakers are also included, is still required. A system based on speech recognition, forced alignment, acoustic-phonetic feature extraction and, for instance, deep learning modeling would enable large-scale automatic ERSUS recognition from linguistically annotated speech data excluding the possibility of the classification of linguistic content instead of emotions.

Acknowledgements Open access funding provided by University of Eastern Finland (UEF) including Kuopio University Hospital. This study was supported by Jenny and Antti Wihuri Foundation (Grant No. 00160426). The author also thanks Emergency Response Centre Administration Finland for cooperation in collecting the research material.

Compliance with ethical standards

Conflict of interest The authors report no conflicts of interest.

References

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program]. Version 6.0.28. http://www.fon.hum.uva.nl/praat/download_linux.html. Accessed 23 May 2017.

Cummings, 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 Communication, 71*, 10–49.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic and linguistic factors. *The Journal of the Acoustical Society of America, 137*(3), 1513–1528.

Demenko, G. (2008). Voice stress extraction. In *Proceedings of Speech Prosody*. Campinas, Brasil. https://pdfs.semanticscholar.org/9d56/57339e1aafb15c81036cfbab636bd8f449ff.pdf. Accessed 24 September 2017.

Demenko, G., & Jastrzebska, M. (2012). Analysis of voice stress in call centers conversations. In *Proceedings of Speech Prosody*. Shanghai, China. https://pdfs.semanticscholar.org/d352/0ac7e52fe17cb6e63f9d5953fb0c7eb17494.pdf. Accessed 24 September 2017.

Farrús, M. (2008). *Fusing prosodic and acoustic information for speaker recognition*. [Dissertation]. Barcelona, Spain: Polytechnic University of Catalonia.

Gałka, J., Grzybowska, J., Igras, M., Jaciów, P., Wajda, K., Witkowski, M., & Ziółko, M. (2015). System supporting speaker identification in emergency call center. In *Proceedings of the Interspeech*. Dresden, Germany. https://pdfs.semanticscholar.org/8b4e/77a70ed4b3587a5e8f9c736d94544762e257.pdf. Accessed 24 September 2017.

Hansen, J. H., & Patil, A. S. (2007). Speech under stress: Analysis, modelling and recognition. In C. Müller (Ed.), *Speaker classification I: Fundamentals, features, and methods* (pp. 108–137). Berlin: Heidelberg.

Harnsberger, J. D., Hollien, H., Martin, C. A., & Hollien, K. A. (2009). Stress and deception in speech: Evaluating layered voice analysis. *Forensic Sciences, 54*, 642–650.

Hautamäki, V., Kinnunen, T., Nosratighods, M., Lee, K.-A., Ma, B., & Li, H. (2010). Approaching human listener accuracy with modern speaker verification. In *Proceedings of the Interspeech*, Makuhari, Japan. pp 1473–1476.

He, L., Lech, M., Maddage, N. C., & Allen, N. B. (2011). Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control, 6*(2), 139–146.

Hill, D. R. (2007). Speaker classification concepts: Past, present and future. In C. Müller (Ed.), *Speaker classification I: Fundamentals, features, and methods* (pp. 21–46). Berlin: Heidelberg.

Hollien, H. (1990). *Acoustics of crime*. New York: Plenum.

Jacob, A. (2017). Modelling speech emotion recognition using logistic regression and decision trees. *International Journal of Speech Technology, 20*(4), 897–905.

Jessen, M. (2006). *Einfluss von stress auf sprache und stimme. Unter besonderer Berücksichtigung polizeidienstlicher Anforderungen*. Idstein: Schulz-Kirchiner Verlag GmbH.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2*(4), 671–711.

Keating, P. A., & Esposito, C. (2006). Linguistic voice quality. In *Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology*. Auckland.

Kirchhübel, C., & Howard, D. M. (2013). Detecting suspicious behaviour using speech: Acoustic correlates of deceptive speech—An exploratory investigation. *Applied Ergonomics, 44*(5), 694–702.

Kirchhübel, C., Howard, D. M., & Stedmon, A. W. (2011). Acoustic correlates of speech when under stress: Research, methods and future directions. *International Journal of Speech Language and the Law, 18*(1), 75–98.

Kreiman, J., & Gerratt, B. R. (2010). Perceptual sensitivity to first harmonic amplitude in the voice source amplitude in the voice source. *The Journal of the Acoustical Society of America, 128*(4), 2085–2089.

Lantz, B. (2013). *Machine learning with R*. Birmingham: Packt Publishing Ltd.

Li, F., Rendall, D., Vasey, P. L., Kinsman, M., Ward-Sutherland, A., & Diano, G. (2016). The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. *Journal of Phonetics, 57*, 59–70.

Meyer, P., Buschermöhle, E., & Fingscheidt, T. (2018). What do classifiers actually learn? a case study on emotion recognition datasets. In *Proceedings of the Interspeech* (pp 262–266). Hyderabad, India.

Milborrow, S. (2017). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 2.1.2. https://CRAN.R-project.org/package=rpart.plot. Accessed 10 November 2017.

Milošević, M., & Đurović, Z. (2015). Challenges in emotion speech recognition. *3rd International Conference on Electrical, Electronic and Computing Engineering. IcETRAN* 2015, Serbia. https://www.researchgate.net/publication/282877098. Accessed 02 February 2018.

Morrison, G. S. (2007). Logistic regression modelling for first and second language perception data. In P. Prieto, J. Mascaró & M.-J. Solé (Eds.), *Segmental and prosodic issues in romance phonology* (pp. 219–236). Amsterdam: John Benjamins Publishing Company.

Morrison, G. S., & Kondaurova, M. V. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America, 126*(5), 2159–2162.

Niemi-Laitinen, T. (1999). *Puhujantunnistus rikostutkinnassa*. [Licentiate thesis]. Helsinki, Finland: University of Helsinki.

Piegorsch, W. W. (2015). *Statistical data analytics: Foundations for data mining, informatics, and knowledge discovery*. Chichester: Wiley.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. Accessed 10 March 2017.

Simpson, A. P. (2012). The first and second harmonics should not be used to measure breathiness in male and female voices. *Journal of Phonetics, 40*(3), 477–490.

Sondhi, S., Khan, M., Vijay, R., Salhan, A. K., & Chouhan, S. (2015). Acoustic analysis of speech under stress. *International Journal of Bioinformatics Research and Applications, 11*(5), 417–432.

Steeneken, H. J., & Hansen, J. H. (1999). Speech under stress conditions: Overview of the effect on speech production and on system performance. *International Conference on Acoustics, Speech, and Signal Processing*; 1999; Phoenix, AZ, USA. IEEE, pp 2079–2082.

Streeter, L. A., Macdonald, N. H., Apple, W., Krauss, R. M., & Galotti, K. M. (1983). Acoustic and perceptual indicators of emotional stress. *The Journal of the Acoustical Society of America, 73*(4), 1354–1360.

Tavi, L. (2017). Acoustic correlates of female speech under stress based on /i/-vowel measurements. *International Journal of Speech Language and the Law, 24*(2), 227–241.

The H2O.ai team (2017). h2o: R Interface for H2O. R package version 3.16.0.2. https://CRAN.R-project.org/package=h2o. Accessed 20 January 2018.

Therneau, T., Atkinson, A., & Ripley, B. (2017). rpart: Recursive partitiong and regression trees. R package version 4.1-10. http://CRAN.R-project.org/package=rpart. Accessed 14 May 2017.

Tracy, E. C., Bainter, S. A., & Satariano, N. P. (2015). Judgments of self-identified gay and heterosexual male speakers: Which phonemes are most salient in determining sexual orientation? *Journal of Phonetics, 52*, 13–25.

Van Lierde, K., van Heule, S., De Ley, S., Mertens, E., & Claeys, S. (2009). Effect of psychological stress on female vocal quality. *Folia Phoniatrica et Logopaedica, 61*(2), 105–111.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th edn.). New York: Springer.

Ververidis, D. & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, *48*(9), 1162–1181.

Vicenik, C. (2017). Phonation-measurements—Praat script. http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html. Accessed 21 May 2017.

Waaramaa, T., Palo, P., & Kankare, E. (2014). Emotions in freely varying and mono-pitched vowels, acoustic and EGG analyses. *Logopedics, Phoniatrics, Vocology, 40*(4), 156–170.

Xu, Y. (2013). ProsodyPro—A Tool for Large-scale Systematic Prosody Analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*; 2013; Aix-en-Provence, France. 2013. p. 7–10. http://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/. Accessed 30 April 2017.

Xu, Y., Kelly, A., & Smillie, C. (2013). Emotional expressions as communicative signals. In S. Hancil & D. Hirst (Eds.), *Prosody and iconity* (pp. 33–61). Amsterdam: John Benjamins Publishing.

Zhou, G., Hansen, J. H., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing, 9*(3), 201–216.