Stress Detection from Speech Using Spectral Slope Measurements

Olympia Simantiraki[1](✉), Giorgos Giannakakis[1], Anastasia Pampouchidou[2], and Manolis Tsiknakis[1,3]

[1] Institute of Computer Science, Foundation for Research and Technology–Hellas (FORTH–ICS), Heraklion, Crete, Greece
{osimantir,tsiknaki}@ics.forth.gr

[2] Le2i Laboratory, University of Burgundy, Le Creusot, France

[3] Department of Informatics Engineering, Technological Educational Institute of Crete, Heraklion, Crete, Greece

Abstract. Automatic detection of emotional stress is an active research domain, which has recently drawn increasing attention, mainly in the fields of computer science, linguistics, and medicine. In this study, stress is automatically detected by employing speech-derived features. Related studies utilize features such as overall intensity, MFCCs, Teager Energy Operator, and pitch. The present study proposes a novel set of features based on the spectral tilt of the glottal source and of the speech signal itself. The proposed features rely on the Probability Density Function of the estimated spectral slopes, and consist of the three most probable slopes from the glottal source, as well as the corresponding three slopes of the speech signal, obtained on a word level. The performance of the proposed method is evaluated on the simulated dataset of the SUSAS corpus, achieving recognition accuracy of 92.06%, when the Random Forests classifier is used.

Keywords: Stress detection · Speech analysis · Glottal source
Fundamental frequency · Spectral tilt
Iterative adaptive inverse filtering · Random forests

1 Introduction

Automatic detection of stress from speech is of great interest, since speech is considered a significant modality in evaluating stress [1]. Although there is no single agreed definition on speech under stress, it can be referred as *"Stress is observable variability in certain speech features due to a combination of unconscious response to stressors and/or conscious control"* [2]. Automatic stress detection concerns several disciplines such as computer science, linguistics, and medicine. The importance of detecting stress automatically lies in the high prevalence of stress in the modern lifestyle [3], enfolding a wide range of applications from cockpit electronics to polygraph testing, health care, robotics, interactive voice

N. Oliver et al. (Eds.): MindCare 2016/Fabulous 2016/IIoT 2015, LNICST 207, pp. 41–50, 2018.
https://doi.org/10.1007/978-3-319-74935-8_5

response systems in call-center applications, and in Human Computer Interfaces (HCIs). Such a system could be very valuable in prioritizing emergency calls in hospital/support line call centers where evaluating the severity of each case may be very critical.

The important contribution of human speech in stress assessment has been proved by several studies (e.g., [4]). Speech is a natural human expression in communication. Some features derived from speech (e.g., glottal spectral slope) are more difficult to be manipulated than others (e.g., pitch, intensity). The elicited speech is affected by the speaker's emotions, since emotions affect the muscle tension, which in turn impacts the vortex-flow interaction pattern in the vocal tract [5]. Although in some cases no noticeable effect is observed, there are many cases where speech alteration under stress is significant and easily perceived. The level of change in speech production depends on the intensity and type of emotion expressed (e.g., anger, fear) and/or the environmental conditions the speaker is located into (e.g., Lombard effect [6]). Other studies focus on stress detection using facial cues derived from eyes, mouth, head behavior and camera based heart activity [7]. Additionally, the combination of audio and visual features has proven profitable for emotion recognition [8].

Overall intensity, Teager Energy Operator (TEO), Mel-Frequency Cepstral Coefficients (MFCCs), and functionals of the fundamental frequency (f_0), are among the most widely adopted features for detecting speech under stress [9]. The influence of stress in speech is evident, as signals derived from high stressed speech result in greater amplitude of the (glottal) waveform, and more asymmetrical glottal pulse as compared to neutral condition. These changes have an impact on the intensity of the spectrum; it is shifted over the spectrum, and concentrated in the higher frequencies. The literature shows that the relative overall energy of the spectrum increases in a stressed condition; however this is not a sufficient indicator on its own. The distribution and/or spectral tilt of the spectrum's energy has also to be considered, as suggested in [10].

In the present study, metrics based on the spectral tilt are examined, proving the significant role of spectral tilt in discriminating stressed from neutral speech with high accuracy. Feature extraction is performed using the simulated dataset of the SUSAS corpus [11]. In order to validate the reliability of the proposed features, the statistical test Mann–Whitney is used, which revealed a statistically significant difference between the stressed and neutral indicators and the Random Forests are used for the classification.

2 Related Work

Several studies have focused on the detection of emotional stress from speech, highlighting the distinctive differences in phonation between stressed and neutral speech [12]. Feature analysis methods for classification of speech under stress have also been proposed [9]. Most studies use simulated stressed speech data for evaluation [13] with specific utterances usually being isolated for the analysis [12,14]. A notable limitation of the latter is that the results degrade as the test conditions drift from the environmental or experimental conditions of the training data [12].

A review of available literature showed that the most widely studied features for discriminating neutral and stressed speech are: overall intensity, TEO, MFCCs and functionals of the fundamental frequency (mainly standard deviation, mean and variance) [9]. *Shah et al.* [15] employed Discrete Wavelet Transform (DWT) for feature extraction, and Artificial Neural Networks (ANN) for classification, achieving 85% recognition accuracy. In *Godin et al.* [9], 6 glottal features were extracted, while Gaussian Mixture Model (GMM) was used as the classifier, achieving detection accuracy of 69%. *Sondhi et al.* [16] suggested to use the mean pitch and the formants (F1, F2) of the human voice as reliable and non-invasive indicators of emotional stress, since they were the acoustic measures providing the most significant change under stress. Eleven subjects participated in this study, providing answers from a specific set of responses: “yes”, “no”, “haan” (“haan” means “yes” in Hindi language). In [12], TEO based features were extracted and Hidden Markov Model (HMM) was used for the stressed speech classifier. The classification error rate for the stress/neutral speech was 4.7/4.6% for the closed-speaker-set system, and 13.6/4.0% for the open-speaker-set system. *Fernandez* and *Picard* [17] explored the use of a feature set based on subband decompositions and the TEO. The corpus used consisted of 598 short speech utterances collected from four subjects driving in a simulator. The best performance obtained, with the speaker-dependent mixture model, achieved an accuracy of 96.4% on the training set, and of 61.2% on a separate testing set. Also, the authors concluded that the performance of the speaker-independent model degrades with respect to the models trained on individual speakers.

A second group of studies [13], have underlined the spectrum significance for discriminating stressed and neutral speech characteristics. *Shukla et al.* [13] extracted the Relative Formant Peak Displacement (RFD) and MFCCs, as features for neutral and stressed speech separation. The simulated speech data (neutral, angry, sad and Lombard) used for evaluation was in Hindi and Indian language. A HMM was used, and the combination of RFD and MFCC achieved 59.53% accuracy. In [18], articulatory, excitation (pitch, duration, intensity) and cepstral based features were estimated using the SUSAS database and an HMM classifier achieving an accuracy of 80.6%.

A different approach was proposed by *Yao et al.* [14], where physical characteristics of the vocal folds were investigated. A novel metric, namely the Muscle Tension Ratio (MTR) was introduced to identify speech under stress. Vowel instances /a/ were isolated for the analysis and ROC curves were used for evaluating and comparing the classification of MTR with the Spectral Flatness Measure (SFM), a conversational method of stress measurement. Experimental results showed that MTR outperforms the conversational method of stress measurement.

Drawn from the reported review of literature, SUSAS is among the most widely employed databases used in defining metrics for the automatic discrimination of stressed and neutral speech. Due to the overt variances in the speech and glottal spectrum of stressed and unstressed speech, the features selected for the investigation in this work are based on spectrum variance.

3 Dataset

As already mentioned, the speech corpus used in this study is the widely used, simulated dataset of SUSAS [11]. It consists of 9 male speakers, uttering isolated-words (e.g., “break”, “enter”, “change”) in a quiet environment. The recordings used are 70 words per speaker, pronounced in four styles: neutral, angry, loud, and Lombard effect. In this study only these three stress types are used as they are the most common employed stress conditions [5]. Speech samples are separated into two general clusters: (a) unstressed speech cluster consisting of the neutral utterances (b) stressed speech cluster consisting of the angry, loud, and Lombard utterances. The total number of samples is 630 words of unstressed speech and 1890 words of stressed speech. The speech tokens were sampled in a 16-bit A/D converter with 8 kHz sampling rate.

4 Feature Extraction

In the speech production model, the acoustic speech signal is the result of the glottal source signal[1] modulated by a transfer (filter) function, the vocal tract. Equation (1), shows mathematically this convolution process

$$s[n] = g[n] \star v[n] \tag{1}$$

where $s[n]$ is the speech signal derived from the convolution of the impulse response of the vocal tract $v[n]$ with the glottal source excitation signal $g[n]$. When a speaker is under stress, both the vocal folds and the movement of the articulators (vocal tract) are affected. Therefore, for a reliable detection of stressed speech, features based on the glottal source signal and speech characteristics should be taken into consideration. In human speech, stressed syllables are produced with greater vocal effort. If a speaker makes greater vocal effort, the amplitudes of the higher frequencies increase more than that of the lower frequencies [19]. For this reason, the use of the spectral tilt is introduced as a measure of the relative distribution of spectral energy from lower to higher frequencies [20]. The proposed features are computed both for the output-speech signal and for the glottal source signal. Furthermore, the standard deviation of the fundamental frequency is extracted, since it has been proven to be a good feature for the separation of stressed and neutral speech [21]. In this study, the analysis of the extracted features is performed only on the voiced speech areas which are discriminated using the f_0 estimation from the SWIPE algorithm [22].

[1] Glottal source signal is the signal generated at the glottis which could be either periodic pulses or noise.

4.1 Fundamental Frequency

The fundamental frequency (f_0) has been widely used in several studies (e.g., [23]) and has proven to be a good indicator for separating stressed from neutral speech. As already mentioned, in this study, the fundamental frequency is estimated only for the voiced areas using the SWIPE algorithm [22] and its standard deviation was selected as the feature to be used in the subsequent analysis. The pitch is searched within a specific range ([70 450] in Herz) and is estimated every 5 ms. The f_0 has been normalized in the range [0, 1].

4.2 Spectral Slopes

In estimating the spectral slopes, the voiced areas are isolated using the f_0 estimation from the SWIPE algorithm. No estimation is made in the unvoiced areas, and thus these frames are excluded from the analysis. Then, the magnitude spectrum is computed using the Fast Fourier Transform (FFT) with 30 ms window length and 5 ms overlap. The spectral envelope [24] is subsequently estimated with optimal spectral order computed by $\lfloor \frac{f_s}{2f_0} \rfloor$ (where f_s is the sampling frequency and f_0 is the fundamental frequency) normalized in dB. Finally, in order to compute the spectral tilt, a linear regression line is fitted to the spectral envelope of the frame using the least square error method and the slope of the regression line is obtained as a measure for the spectral tilt. The same procedure is repeated for all voiced segments of each word. The probability density function (PDF) is computed for each word for the bag-of-slopes extracted (one slope per voiced frame). Then, the three most probable slopes of PDF histogram are used as features for the classification. The bin width in the histogram is 0.017 (this value corresponds to $\pi/180$ rad).

Glottal Source Signal. During stressed phonation, a combination of changes in sub-glottal air pressure can lead to irregular shape of the glottal pulses [25]. In Fig. 1, the glottal pulses for the stressed speech (angry) and neutral speech are depicted. In order to estimate the glottal source signal, we employed the Iterative Adaptive Inverse Filtering (IAIF) method [26] using linear prediction for the estimation of the vocal tract response. The IAIF removes the vocal tract effects in an iterative manner in order to obtain an accurate estimation of the glottal source signal. The observed differences are also reflected on the glottal spectrum, resulting in increased energy at higher frequency areas. The mean glottal spectrums and the corresponding spectral tilts, computed on the same voiced part for each condition, i.e. stressed and neutral, are illustrated in Fig. 2(a), in which differences can be clearly observed. Based on the observation regarding the characteristics of the spectrum, feature extraction based on the spectral tilt of the glottal source spectrum is proposed herein. In Fig. 2(b), the PDF of the spectral tilt for the voiced tokens of the spoken word /NAV/, for both stressed and unstressed speech styles are shown. The three highest peaks of the PDF curves for each word are selected for the classification. Additionally, in Fig. 2(b),

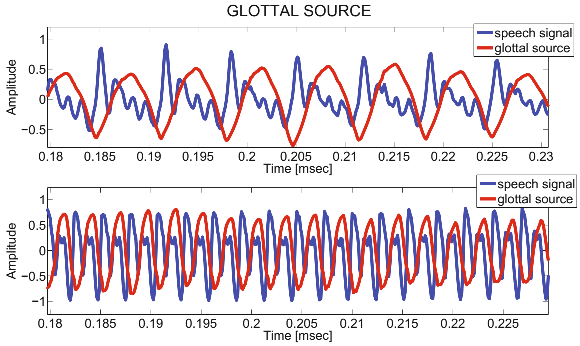
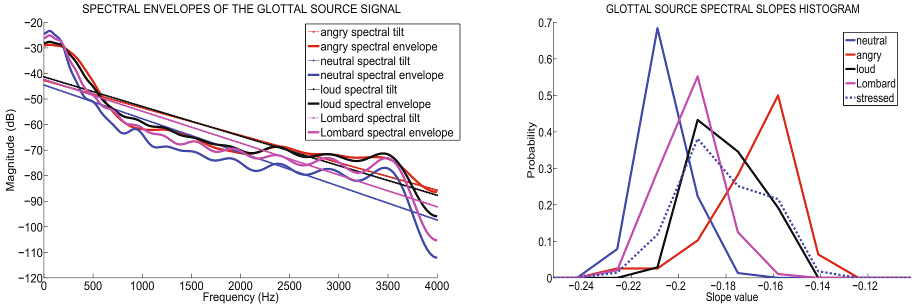


Fig. 1. *Upper panel*: a neutral speech segment (*blue line*) and its corresponding glottal source (*red line*). *Lower panel*: a stressed (angry) speech signal (*blue line*) and its corresponding glottal source (*red line*). Token /a/ is uttered by a male speaker in both panels, isolated from the word /NAV/. (Color figure online)

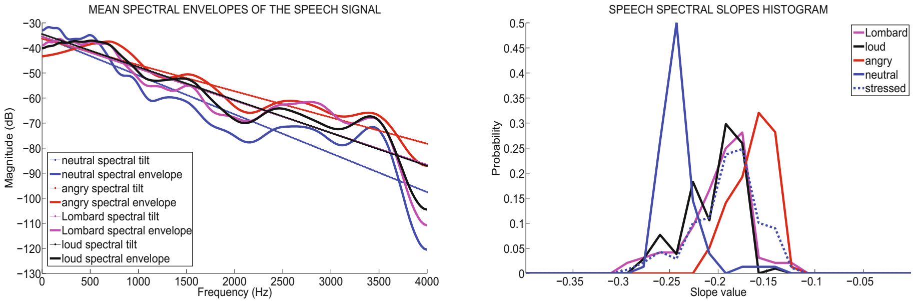


(a) (b)

Fig. 2. (a) *Solid lines*: mean values of the glottal source spectral envelopes. *Dashed lines*: corresponding spectral tilts. (b) Probability Density Functions of the glottal spectral slopes. All information is extracted from the voiced tokens isolated from the word /NAV/ uttered by a male speaker, for neutral speech (*blue solid line*), 3 different stressed speech styles (*red, black, magenta solid lines*), and the combined stressed speech styles (*blue dashed line*). (Color figure online)

we observe that the PDF curve of the stressed speech is to the right of the neutral curve, which means that the glottal spectrum tilt of the stressed speech is greater than that of the neutral. Most commonly, if the glottal waveshape is less smooth, it will have a stronger harmonic structure and a spectral slope closer to zero [9], as opposed to neutral speech, for which the spectral slope for voiced frames is typically negative.

Speech Signal. A stressed speaker utilizes a more pronounced and loud voice, changing the shape of the vocal tract formants. More precisely, in a condition of stress the location of the formants and their bandwidth are different than in neutral speech [21,25]. Also the filter function is excited by the source spectrum, resulting in a tilt in the overall spectrum. The stress affects higher frequency regions more than that of the lower frequency regions. Therefore, under stress conditions, the spectral tilt increases with respect to neutral condition (Fig. 3(a)). As the spectrum of the speech signal could enhance the glottal spectrum information, the speech spectral slope is also prominent for characterization and classification of stressed speech. The three highest peaks in the PDF curve (Fig. 3(b)) for each word are used as additional features for the classification.



(a) (b)

Fig. 3. (a) *Solid lines*: mean values of the speech signal spectral envelopes. *Dashed lines*: corresponding spectral tilts. (b) Probability Density Functions of the speech signal spectral slopes. All information is extracted from the voiced tokens isolated from the word /NAV/ uttered by a male subject, for neutral speech (*blue solid line*), 3 different stressed speech styles (*red, black, magenta solid lines*), and the combined stressed speech styles (*blue dashed line*). (Color figure online)

5 Statistical and Classification Analysis

The statistically significant difference between the stressed and neutral PDF curves, from which the spectrum based features have been extracted, is quantified with the statistical test Mann–Whitney. Figures 2(b) and 3(b), illustrate this significant difference of the distributions resulting from the spectral tilt of stressed and neutral speech. Two separate Mann–Whitney statistical tests are performed. In the first test, the glottal and speech spectral slopes of the neutral style and the corresponding spectral slopes of all stressed styles are compared using the statistical test for each speaker individually. The individual difference of the distributions is statistically significant ($p < 0.01$). In the second test, the same statistical test is performed between the neutral and stressed spectral slopes (glottal and speech slopes separately) for all nine speakers yielding statistically significant differentiation between the distributions ($p < 0.01$). Based on the reported findings, the spectral slopes can be characterized as significant informative indicators for distinguishing neutral from stressed speech.

The Random Forests (RF) algorithm [27] is applied in order to evaluate the performance of the proposed features (three glottal spectral slopes, three speech spectral slopes, and the standard deviation of the f_0), for the discrimination between stressed and unstressed speech. The RF is an aggregation of decision trees algorithm. This method is considered as a general technique of decision trees, and is an ensemble learning method for classification. The main advantage of the Random Forests algorithm, is that it corrects the decision trees' tendency of over-fitting to their training set. More specifically, a text-independent classification is performed using all of the 2520 recordings from the SUSAS dataset. For the evaluation, we apply the Random Forests classification technique, with 1000 decision trees in the ensemble, while the rest of their tuning hyper-parameters are set to the default values in MATLAB. A 10 times repeated 10-fold cross-validation is performed, achieving a classification accuracy of 92.06%. This performance is probably an underestimation due to the use of only one specific classifier [28]. In Fig. 4, an estimation of the importance of each feature for the accuracy of the classifier is depicted. It is evident that the contribution of three of the proposed features in the system's learning is greater than that of the f_0's.

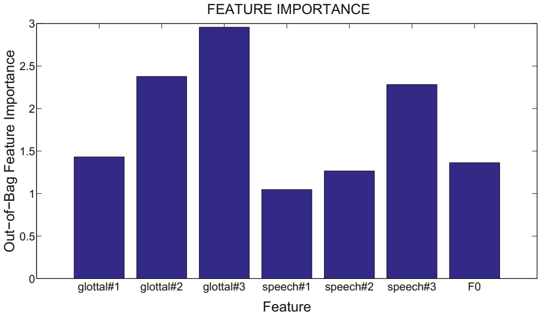


Fig. 4. Proposed features are denoted with the labels *glottal#1, glottal#2, glottal#3, speech#1, speech#2, speech#3 (glottal#n are the 3 most probable glottal spectral slopes provided by the PDF histogram and speech#n the corresponding speech spectral slopes)* and with $F0$ the f_0 std.

6 Conclusion

This study focuses on separating the neutral from stressed speech, using features estimated on the frequency domain. More specifically, the features used are the three most probable slopes of the speech spectrum provided by the PDF histogram, the corresponding three slopes of the glottal spectrum, and the standard deviation of the f_0. These features are extracted for the voiced segments of each word in the dataset and not for specific tokens. The main advantage of the proposed features is that they cannot be easily manipulated for concealing stress, in contrast to the most commonly used features in similar studies, e.g., overall intensity and pitch, which can be voluntarily controlled. Therefore they form the basis for a more objective estimation method. Furthermore, it is not sufficient to deal only with the overall intensity, since a shift of the intensity over the spectrum is observed when more effort is made for the speech production. This shift is apparent on the spectral tilt. In this study, the classification performed using the metrics based on spectral tilt and f_0 was text-independent achieving an accuracy score of 92.06%. As a result, our main conclusion is that these metrics are capable of distinguishing the two speaking styles with great success.

Future work will focus on extended data analysis by testing different classifiers, on applying the proposed features on datasets with stressed and neutral recordings under real conditions and on evaluating the detection accuracy by combining these speech features with visual features, in an analysis approach similar to the one in [29].

Acknowledgments. The authors acknowledge support from the iManageCancer EU project under contract H2020-PHC-26-2014 No.643529.

References

1. Sharma, N., Gedeon, T.: Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Comput. Methods Programs Biomed. **108**(3), 1287–1301 (2012)
2. Murray, I.R., Baber, C., South, A.: Towards a definition and working model of stress and its effects on speech. Speech Commun. **20**(1), 3–12 (1996)
3. Selye, H.: The Stress of Life. McGraw-Hill, New York (1956)
4. Lefter, I., Rothkrantz, L.J., Van Leeuwen, D.A., Wiggers, P.: Automatic stress detection in emergency (telephone) calls. Int. J. Intell. Defence Support Syst. **4**(2), 148–168 (2011)
5. Zhou, G.J., Hansen, J.H.L., Kaiser, J.F.: Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. **9**(3), 201–216 (2001)
6. Garnier, M., Henrich, N.: Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? Comput. Speech Lang. **28**(2), 580–597 (2014)
7. Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P.G., Marias, K., Tsiknakis, M.: Stress and anxiety detection using facial cues from videos. Biomed. Signal Process. Control **31**, 89–101 (2017)
8. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 39–58 (2009)
9. Godin, K.W., Hasan, T., Hansen, J.H.: Glottal waveform analysis of physical task stress speech. In: INTERSPEECH, pp. 1648–1651 (2012)
10. Sluijter, A.M., Van Heuven, V.J.: Spectral balance as an acoustic correlate of linguistic stress. J. Acoust. Soc. Am. **100**(4), 2471–2485 (1996)
11. Hansen, J.H., Bou-Ghazale, S.E., Sarikaya, R., Pellom, B.: Getting started with SUSAS: a speech under simulated and actual stress database. In: Eurospeech, vol. 97(4), pp. 1743–46 (1997)

12. Hansen, J.H., Kim, W., Rahurkar, M., Ruzanski, E., Meyerhoff, J.: Robust emotional stressed speech detection using weighted frequency subbands. EURASIP J. Adv. Signal Process. **2011**(1), 1–10 (2011)
13. Shukla, S., Dandapat, S., Prasanna, S.R.M.: Spectral slope based analysis and classification of stressed speech. Int. J. Speech Technol. **14**(3), 245–258 (2011)
14. Yao, X., Jitsuhiro, T., Miyajima, C., Kitaoka, N., Takeda, K.: Physical characteristics of vocal folds during speech under stress. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4609–4612 (2012)
15. Shah, F., Sukumar, R., Anto, B.: Automatic Stress Detection from Speech by Using Discrete Wavelet Transforms (2009)
16. Sondhi, S., Khan, M., Vijay, R., Salhan, A.K.: Vocal indicators of emotional stress. Int. J. Comput. Appl. **122**(15), 38–43 (2015)
17. Fernandez, R., Rosalind, W.P.: Modeling drivers speech under stress. Speech Commun. **40**(1), 145–159 (2003)
18. Womak, B.D., Hansen, J.H.: Improved speech recognition via speaker stress directed classification. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 1, pp. 53–56 (1996)
19. Eriksson, A., Traunmüller, H.: Perception of vocal effort and distance from the speaker on the basis of vowel utterances. Percept. Psychophysics **64**(1), 131–139 (2002)
20. Tartter, V.C., Gomes, H., Litwin, E.: Some acoustic effects of listening to noise on speech production. J. Acoust. Soc. Am. **94**(4), 2437–2440 (1993)
21. Sigmund, M.: Introducing the database ExamStress for speech under stress. In: Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG, pp. 290–293. IEEE (2006)
22. Camacho, A.: SWIPE: A sawtooth waveform inspired pitch estimator for speech and music. Doctoral dissertation, University of Florida (2007)
23. Protopapas, A., Lieberman, P.: Fundamental frequency of phonation and perceived emotional stress. J. Acoust. Soc. Am. **101**(4), 2267–2277 (1997)
24. Röbel, A., Rodet, X.: Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In: International Conference on Digital Audio Effects, pp. 30–35 (2005)
25. Hansen, J.H.L., Patil, S.: Speech under stress: analysis, modeling and recognition. In: Müller, C. (ed.) Speaker Classification I. LNCS (LNAI), vol. 4343, pp. 108–137. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74200-5_6
26. Alku, P.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. **11**(2–3), 109–118 (1992)
27. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2008)
28. Tsamardinos, I., Rakhshani, A., Lagani, V.: Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. In: Likas, A., Blekas, K., Kalles, D. (eds.) SETN 2014. LNCS (LNAI), vol. 8445, pp. 1–14. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07064-3_1
29. Pampouchidou, A., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., Giannakakis, G., Meriaudeau, F., Simos, P., Marias, K., Yang, F., Tsiknakis, M.: Depression assessment by fusing high and low level features from audio, video, and text. In: The 6th Audio/Visual Emotion Challenge and Workshop. ACM-Multimedia (2016)