Speech Features for Discriminating Stress Using Branch and Bound Wrapper Search

Mariana Julião[1]([✉]), Jorge Silva[1], Ana Aguiar[1], Helena Moniz[2,3],
and Fernando Batista[2,4]

[1] Instituto de Telecomunicações,
Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
{mjuliao,ana.aguiar}@fe.up.pt, up201007483@alunos.dcc.fc.up.pt
http://www.it.pt
[2] INESC-ID, Lisboa, Portugal
{helenam,fmmb}@l2f.inesc-id.pt
[3] FLUL/CLUL, Universidade de Lisboa, Lisboa, Portugal
[4] ISCTE - Instituto Universitário de Lisboa, Lisboa, Portugal

Abstract. Stress detection from speech is a less explored field than Automatic Emotion Recognition and it is still not clear which features are better stress discriminants. The project VOCE aims at doing speech classification as stressed or not-stressed in real-time, using acoustic-prosodic features only. We therefore look for the best discriminating feature subsets from a set of 6125 features extracted with openSMILE toolkit plus 160 Teager Energy Operator (TEO) features. We use a Mutual Information (MI) filter and a branch and bound wrapper heuristic with an SVM classifier to perform feature selection. Since many feature sets are selected, we analyse them in terms of chosen features and classifier performance concerning also true positive and false positive rates. The results show that the best feature types for our application case are Audio Spectral, MFCC, PCM and TEO. We reached results as high as 70.4 % for generalisation accuracy.

Keywords: Stress · Emotion recognition · Ecological data · Feature sets · Feature selection

1 Introduction

The motivations for detecting stress from speech range from it being a non-intrusive way to detect stress, to ranking emergency calls [7], or improve speech recognition systems, since it is known that environmentally induced stress leads to fails on speech recognition systems [13]. Public Speaking is said to be "the most common adult phobia" [18], showing the relevance of a tool to improve public speaking. In VOCE[1], we target developing such a tool, by developing algorithms to identify emotional stress from live speech. In particular, VOCE

---

[1] http://paginas.fe.up.pt/~voce.

corpus comes mainly from public speaking events that occur within academic context, like presentations of coursework or research seminars. The envisioned coaching application requires detecting emotional stress in live speech in near real time, to give the user timely feedback, which requires adapting the computational costs to the limited memory and computational resources to use. Decreasing the number of features used for classification reduces the amount of data to collect, the amount of features to be extracted and the complexity of the classifier, impacting a reduction in the memory and computational resources used. Additionally, feature selection can increase the classifier's accuracy [12]. Thus, in this paper, we focus on identifying these reduced feature sets based on their performance as stress discriminators.

In this work, we start from the fusion of two feature sets: the group of features extracted using the openSMILE toolkit [25], and the group of TEO-based features, to be detailed on Sect. 4.2. We filter these feature sets with Mutual Information (MI) and then use a branch-and-bound wrapper to explore the space of possible feature sets. Finally, we analyse the best feature sets chosen on various branches for the most frequently chosen feature categories.

2   Related Work

The importance of suprasegmental acoustic phenomena that can be taken as global emotion features is highlighted in [28], like "hyper-clear speech, pauses inside words, syllable lengthening, off-talk, disfluency cues, inspiration, expiration, mouth noise, laughter, crying, unintelligible voice". These features have been mainly annotated by hand, and automatic extraction is not straightforward, though possible in some cases.

Stress recognition from speech is a specific case of emotion recognition. The Fundamental Frequency, F0, is the most consensual feature for stress discrimination [8,14,22,31], but several metrics for energy and formant changes have been proposed, often represented by Mel-Frequency Cepstral Coefficients (MFCCs) [7,21,31]. Frequency and amplitude perturbations – Jitter and Shimmer –, and other measures of voice quality, like Noise to Harmonics Ratio and Subharmonics to Harmonics Ratio [26,28] have also been used. Teager Energy Operator-based features have also shown to perform well in speech under stress [31], and we shall look at them in detail in this work.

TEO-based features have been shown to increase recognition robustness with car noise [10,15]. In [17], TEO-based features reached the best performance for stressed speech discrimination outdoor, but not indoor. They also have been used to do voiced-unvoiced classification [19]. In the latter work, the advantages of TEO are enunciated: because only three samples are needed for the energy computation at each time instant, it is nearly instantaneous. Therefore, this time resolution allows to capture energy fluctuation, and also a robust AM-FM estimation in noisy environment. [6] uses Teager Energy Operator in the development of a system for hypernasal speech detection. In this work, we shall look into the discrimination power of TEO-based speech features for stress discrimination in public speaking.

3   Speech Corpus and Data Annotation

The VOCE corpus [2] currently consists of 38 raw recordings from voluntaries aged 19 to 49. Data is recorded in an ecological environment, concretely during academic presentations[2]. Speech was automatically segmented into utterances, according to a process described in [5].

Annotation into stressed or neutral classes was performed per speaker, based on the mean heart rate [4]. Utterances on the third quartile of mean heart rates for that speaker are annotated as stressed, while the remaining ones are annotated as neutral.

Using an ecologically collected corpus imposes an unavoidable trade-off between the quality of the recording and the spontaneity of the speaker. Higher quality of the recording not only allows for more reliable feature extraction, in general, but also impacts the performance of the segmentation algorithms we use to split the speech into sentence-like units – utterances –, and to do text transcription, necessary for the extraction of TEO features. For these reasons, we chose only 21 raw recordings for this work.

For these speakers, 1457 valid utterances were obtained[3]. The set of utterances is divided into 15 speakers (507 utterances) for training and 6 speakers

Table 1. Dataset demographic data. PSE: Public Speaking Experience, $1-5$: 1 - little experience, 5 - large experience.

Train set				Test set			
Age	Gender	PSE	#Utts	Age	Gender	PSE	#Utts
26	male	2	56	24	male	3	97
22	male	2	39	19	male	2	61
24	male	3	36	19	male	3	86
21	male	3	38	19	female	3	64
22	male	3	32	23	female	4	71
22	male	3	25	19	female	3	63
25	male	2	54				
19	male	3	12				
21	male	3	22				
21	female	3	51				
24	female	5	27				
22	female	2	37				
21	female	3	32				
21	female	3	18				
19	female	3	28				

---

[2] Please refer to [3] for details on the collection methodology.
[3] Remaining utterances after discarding 94 utterances with length of less than 1 s or more than 25 s.

(442 utterances) for testing. Since the number of stressed utterances corresponds to approximately 1/4 of the total, we randomly down-sampled the train data in order to balance the two classes, which led to the mentioned 507 utterances. During feature selection, the classifier was trained on 354 utterances and tested on 153 utterances. These utterances belonged to the train set. Table 1 characterises the dataset concerning age, gender, public speaking experience, and the number of utterances considered[4].

We performed outlier detection on each feature using the Hampel identifier [20] with t = 10. The outliers were then replaced by the mean value of the feature excluding outliers, and feature values were scaled to the interval [0,1].

4   Methodology

Figures 1(a) and (b) illustrate the workflow for speech segmentation and feature selection, respectively. In this work, we start from the fusion of two feature sets: the group of features extracted using the openSMILE toolkit [25], and the group of TEO-based features, to be detailed on Sect. 4.2. We filter these feature sets with Mutual Information and then use a branch-and-bound wrapper to explore the space of possible feature sets. We then analyse the best feature sets chosen on various branches for the most frequently chosen feature categories.
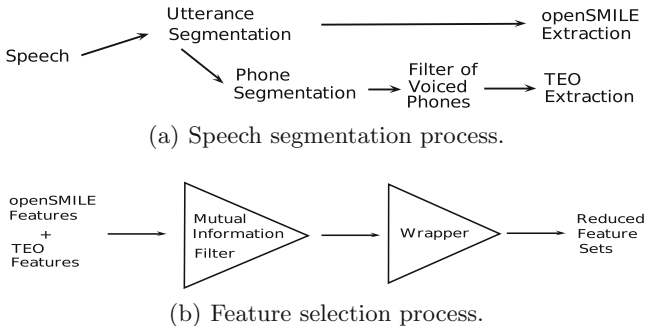
Speech → Utterance Segmentation → openSMILE Extraction

Phone Segmentation → Filter of Voiced Phones → TEO Extraction

(a) Speech segmentation process.

openSMILE Features + TEO Features → Mutual Information Filter → Wrapper → Reduced Feature Sets

(b) Feature selection process.

Fig. 1. Workflow for the speech segmentation and the feature selection process.

4.1   Acoustic-Prosodic Features

OpenSMILE extracts a set of 128 low-level features at the frame level from the speech signal, known as low-level descriptors (LLD) [11]. Statistical functionals are then applied over the LLD in order to compute values for longer segments, providing a total of 6125 features at the segment level [25]. These features and their extraction processes are described in [9,24].

---

[4] Please note that the stated number of utterances on the train set corresponds to the one actually used after discarding a part of the neutral utterances, and not to the number of utterances in the natural set.

The openSMILE toolkit is capable of extracting a very wide range of acoustic-prosodic features and has been applied with success in a number of paralinguistic classification tasks [23]. It has been used in the scope of this study to extract a feature vector containing 6125 speech features, by applying segment-level statistics (means, moments, distances) over a set of energy, spectral and voicing related frame-level features.

4.2   Teager Energy Operator Features

The following TEO-Based features were extracted: Normalized TEO autocorrelation envelope and Critical Band Based TEO Autocorrelation Envelope as in [31]. The literature where Normalized TEO Autocorrelation Envelope and Critical Band Based TEO Autocorrelation are presented targets the feature extraction for small voiced parts usually called "tokens" [31]. To work equivalently, we did a phone recognition with the delimitation of each phone [1] and used only voiced sounds. These correspond to phones represented by the portuguese SAMPA symbols 'i', 'e', 'E', 'a', '6', 'O', 'o', 'u', '@', 'i~', 'e~', '6~', 'o~', 'u~', 'aw', 'aj', '6~j~', 'v', 'z', 'Z', 'b', 'd', 'g', 'm', 'n', 'J', 'r', 'R', 'l', 'L' [29, Chap. IV.B].

These features are extracted per frame. The length of each frame is about 10ms, depending on the feature to extract. Each phone usually contains many frames and each utterance has normally many phones. Therefore, since we want to have values per utterance, we consider each feature extracted for all phones and apply statistics to it. These statistics are: mean, standard deviation, skewness, kurtosis, first quartile, median, third quartile, and inter-quartile range. This process is also illustrated in Fig. 1(a). The first two columns in Table 2 summarise the feature types considered in this work[5].

5   Searching for the Best Feature Sets

As already stated, we apply one filter to reduce the dimensionality from initially 6285 functional (OS) plus TEO features before applying the wrapper with a Support Vector Machine (SVM) classifier with radial basis function kernel and C=100[6], using *python* library *scikit-learn*.

5.1   Filter: Mutual Information

There are several metrics and algorithms to compute the relevance of features on a dataset, and the choice of this metric may hugely impact the final subset of features. However, since there is a lack of a priori knowledge about filter metric adequacy to specific datasets [30], we based our choice on the work of Sun and

---

[5] The generic designation "type" is the result of aggregating Low Level Descriptor features with their derived functionals (e.g., quartiles, percentiles, means, maxima, minima). This procedure is, in our perspective, a way to better group and interpret the performance of the features.

[6] This value was found empirically to produce the best classification results.

Table 2. Feature Types: Id, Name, Number of features of each type selected for MI, Number of features of each type chosen for the Best Sets: T.A.1, T.A.2, G.A., Se., Sp., and Comb.

Id	Type	MI	T.A.1	T.A.2	G.A.	Se.	Sp.	Comb.
1	F0final[a]	10	0	0	0	0	0	0
2	TEO[b]	17	3	2	1	3	2	3
3	audSpec_Rfilt[b]	187	8	8	6	7	6	6
4	audspec[a]	6	0	0	0	0	0	0
5	audspecRasta[a]	4	0	0	0	0	0	0
6	jitterDDP	6	0	0	1	0	0	0
7	jitterLocal[a]	7	0	0	0	0	0	0
8	logHNR	8	1	0	0	0	1	1
9	mfcc_sma[b]	119	7	8	4	2	3	4
10	pcm_Mag_fband	17	0	0	1	2	1	0
11	pcm_Mag_harmonicity	14	1	0	0	0	0	0
12	pcm_Mag_psySharpness[b]	6	3	2	1	1	1	2
13	pcm_Mag_spectralEntropy	6	0	1	0	0	0	1
14	pcm_Mag_spectralFlux	8	0	1	0	0	0	0
15	pcm_Mag_spectralKurtosis[a]	7	0	0	0	0	0	0
16	pcm_Mag_spectralRollOff	22	1	0	0	0	0	0
17	pcm_Mag_spectralSkewness[a]	1	0	0	0	0	0	0
18	pcm_Mag_spectralSlope	6	0	1	1	0	0	0
19	pcm_Mag_spectralVariance[a]	10	0	0	0	0	0	0
20	pcm_RMSenergy	6	1	1	1	1	1	0
21	pcm_zcr	8	0	0	0	1	0	0
22	shimmerLocal[a]	8	0	0	0	0	0	0
23	voicingFinalUnclipped[a]	4	0	0	0	0	0	0

[a] Type not selected by the best sets
[b] Type always selected by the best sets

Li et al. [27], which showed good results in terms of classification for Mutual Information (MI), a metric that measures the mutual dependence between two random variables.

Since MI is based on the probability distribution of discrete variables and our features have continuous values, we had to define a binning. We (1) defined five binning possibilities: 50, 100, 250, 500 or 1000 bins; (2) computed MI for each feature and each binarisation possibility; (3) kept features for which the MI value belonged to the higher quartile for all binarisation options. Their distribution per feature type corresponds to the third column in Table 2.

5.2   Wrapper

Feature selection has been widely studied and, as result, a large number of algorithms have been proposed. These algorithms can be categorized into three groups: filter, wrapper and embedded [16]. Wrapper algorithms find the final solution using a learning algorithm as part of the evaluation criteria. The main idea of these methods is to use the learning algorithm as a black-box to guide the search for the optimal solution. The learning is applied to every candidate solution and the goodness of the subset is given according to the performance of the learning algorithm. Due to the learning algorithm being directly used on the process of selecting features, these methods tend to find better solutions. Nonetheless, the final solution only applies for the selected learning algorithm, since using a different one will most likely result on a different final solution. These methods have higher computational cost as they require training and classifying data for each candidate solution.

We designed a branch and bound wrapper to search the space of feature sets obtained from the MI filter for the combination of features that deliver the best classifier performance. This wrapper starts by searching all combinations of sets up to 10 features, keeping all that are within 1.5 % accuracy of the best solution found so far. Larger feature sets are obtained by expanding the previously kept solutions with blocks of features not yet in the sets. Every time a feature subset is tested with a classification algorithm, a score is produced, which is the accuracy, in this case. Subsets are kept and expanded if the expansion improves the previous accuracy. This search runs until the work list of feature sets with new combinations empties. This wrapper provides a better exploration of the feature set space than traditional forward and backward wrappers. Since the search space for our wrapper is much bigger than for most wrapper methods, we used parallel programming techniques to improve the throughput of the algorithm, using *python*'s*Multiprocessing* package.

6   Results

The Mutual Information filter selected 487 features, distributed into types as described in the third column of Table 2. After choosing the best 280 feature sets with training accuracies below 85 % from 20 processors, we looked at their distribution by feature types, which is on Fig. 2.

Among these 280 feature sets we looked for the ones having the best scores in each of the considered metrics[7]: Train Accuracy, Generalisation Accuracy, Sensitivity (Se), Specificity (Sp)[8], and a Combined Metric defined as

$$\mathrm{CombinedMetric} = \frac{(\mathrm{Se} + \mathrm{Sp})}{2} - |\mathrm{Sp} - \mathrm{Se}|. \tag{1}$$

---

[7] Generalisation Accuracy, Sensitivity and Specificity are computed on the test set.

[8] Being TP - number of True Positives, TN - number of True Negatives, FP - number of False Positives, FN - number of False Negatives, Sensitivity=$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and Specificity=$\frac{\mathrm{TN}}{\mathrm{TN+FP}}$.

The need for this metric follows from the fact that it is our goal not only to have a good generalisation accuracy, but also to have high sensitivity and high specificity at the same time. This is relevant since, as we have an imbalanced test set, with much more neutral utterances than stressed utterances, it can happen that high generalisation results are due to high values of true positives, while true negatives are neglected – which is the kind of scenario we want to avoid. On Table 3, each line corresponds to the best feature subset for which the metric specified in the first column was found to be maximum. The two last lines correspond to baseline results, meaning the classification for the whole set of features and for the set of MI filtered features.

Columns T.A.1, T.A.2, G.A., Se., Sp., and Comb, in Table 2, correspond to the best feature sets, according to each of these metrics, as exposed in Table 3. Each of the Columns in Table 2 signs the number of features of each type (each line corresponds to a feature type).

Table 3. Metrics for the Best Subsets as percentage

Set	Train Acc.	Gen. Acc.	Sens.	Spec.	Comb.	# features
Train Acc	84.97	61.76	59.81	62.39	58.53	25
Train Acc	84.97	62.22	52.34	65.37	45.82	24
Gen. Acc	81.70	70.36	33.64	82.09	9.42	16
Sensitivity	81.70	59.28	71.96	55.22	46.85	17
Specificity	81.05	70.14	31.78	82.39	6.47	15
Combined	81.70	64.03	61.68	64.78	60.14	17
Complete	—	63.12	50.47	67.16	42.13	6285
MI	—	60.86	45.79	65.67	35.85	487

Table 3 bears the following information:

– The sets of best train accuracy do not correspond to the ones with best generalisation accuracy. Actually, these have the second worse generalisation results among these sets.
– The set of best generalisation accuracy, as well as the set of best specificity, although having very good generalisation accuracies have very low sensitivities. This is the kind of imbalance we want to avoid.
– The same train accuracy can have sets of very different quality. We see that for train accuracy 81.70 % we have the best generalisation accuracy, the best sensitivity and the best combined metric. Looking at the other columns in the table we see that only the line for Combined Metric has acceptable results in sensitivity and specificity.
– *These best reduced sets often achieve better results than both the complete set and the filtered set, having much smaller sets, which is very good for the envisioned real-time public speaking coaching application.*
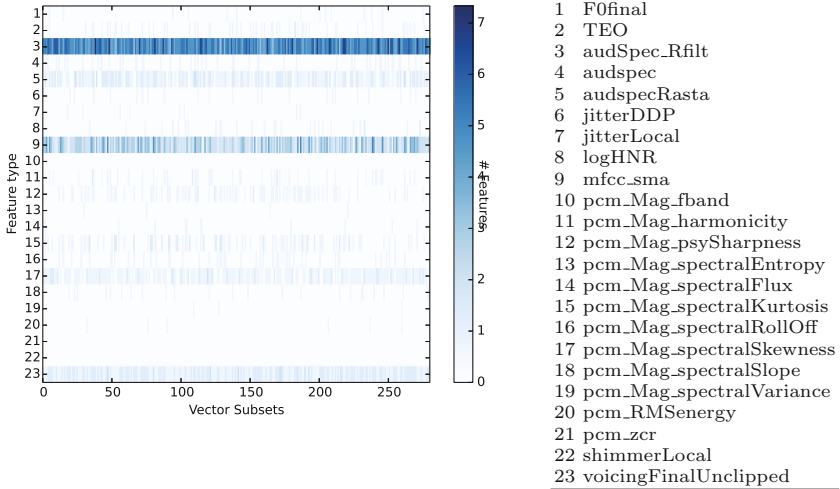
1	F0final
2	TEO
3	audSpec_Rfilt
4	audspec
5	audspecRasta
6	jitterDDP
7	jitterLocal
8	logHNR
9	mfcc_sma
10	pcm_Mag_fband
11	pcm_Mag_harmonicity
12	pcm_Mag_psySharpness
13	pcm_Mag_spectralEntropy
14	pcm_Mag_spectralFlux
15	pcm_Mag_spectralKurtosis
16	pcm_Mag_spectralRollOff
17	pcm_Mag_spectralSkewness
18	pcm_Mag_spectralSlope
19	pcm_Mag_spectralVariance
20	pcm_RMSenergy
21	pcm_zcr
22	shimmerLocal
23	voicingFinalUnclipped

Fig. 2. Heatmap for feature type frequencies on each subset.

7   Discussion

The set of features selected by the Mutual Information filter are, *grosso modo*, the ones reported in the literature for other languages (e.g., [14,32]). Those encompass pitch information, mostly final movements of pitch, audio spectral differences, voice quality features (jitter, shimmer, and harmonics-to-noise-ratio) and TEO features, the latter usually described as very robust across gender and languages. As for PCMs and MFCCs, these features are very transversal in speech processing tasks and highly informative for a wide range of tasks, not surprising, thus, for stress detection as well. The features selected by Mutual Information filter give us a more complete characterization of stress predictors. From these set the ones that are systematically chosen in the best features sets using the wrapper are mostly TEO, MFCCs and audio spectral differences. TEO and MFCCs features are also reported by [32], for English and Mandarin, as the most informative ones, even more than pitch itself.

8   Conclusions

We have used a corpus of ecologically collected speech to search for the best speech features that discriminate stress. Starting from 6125 features extracted with openSMILE toolkit and 160 Teager Energy features, we used a Mutual Information filter to obtain a reduced subset for stress detection. Next, we searched for the best feature set using a branch and bound wrapper with SVM classifiers.

Our results provide further evidence that the features resulting from the Mutual Information filtering process are robust for stress detection tasks, independently of the language, and highlight the importance of voice quality features

for stress prediction, mostly high jitter and shimmer and low harmonics to noise ratio, parameters typically associated with creaky voice.

Our best result compares well with work done by [10,32], although direct comparisons are hard to establish due to different corpora, segmentations, and metrics used in the studies.

References

1. Abad, A., Astudillo, R.F., Trancoso, I.: The L2F spoken web search system for mediaeval 2013. In: Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 18–19 October 2013 (2013)
2. Aguiar, A., Kaiseler, M., Meinedo, H., Almeida, P., Cunha, M., Silva, J.: VOCE corpus: ecologically collected speech annotated with physiological and psychological stress assessments. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). European Language Resources Association (ELRA), Reykjavik (2014)
3. Aguiar, A.C., Kaiseler, M., Meinedo, H., Abrudan, T.E., Almeida, P.R.: Speech stress assessment using physiological and psychological measures. In: Mattern, F., Santini, S., Canny, J.F., Langheinrich, M., Rekimoto, J. (eds.) UbiComp (Adjunct Publication), pp. 921–930. ACM (2013)
4. Allen, M.T., Boquet, A.J., Shelley, K.S.: Cluster analyses of cardiovascular responsivity to three laboratory stressors. Psychosom. Med. **53**(3), 272–288 (1991)
5. Batista, F., Moniz, H., Trancoso, I., Mamede, N.J.: Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Trans. Audio Speech Lang. Process. **20**(2), 474–485 (2012)
6. Cairns, D.A., Hansen, J.H.L., Kaiser, J.F.: Recent advances in hypernasal speech detection using the nonlinear teager energy operator. In: ICSLP 1996, p. 1 (1996)
7. Demenko, G.: Voice stress extraction. In: Proceedings of the Speech Prosody 2008 Conference (2008)
8. Demenko, G., Jastrzebska, M.: Analysis of voice stress in call centers conversations. In: Proceedings of Speech Prosody, 6th International Conference, Shanghai, China (2012)
9. Eyben, F., Wllmer, M., Schuller, B.: openSMILE: the munich versatile and fast open-source audio feature extractor. In: Bimbo, A.D., Chang, S.F., Smeulders, A.W.M. (eds.) ACM Multimedia, pp. 1459–1462. ACM (2010)
10. Fernandez, R., Picard, R.W.: Modeling drivers' speech under stress. Speech Commun. **40**(1–2), 145–159 (2003)
11. Ferreira, J., Meinedo, H.: VOCE project stress feature survey technical report 2. Technical report, L2F, Inesc-ID, Lisboa, Portugal, November 2013
12. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)

13. Hansen, J.H., Bou-Ghazale, S.E., Sarikaya, R., Pellom, B.: Getting started with the susas: A speech under simulated and actual stress database. Technical Report: RSPL-98-10 (1998)
14. Hansen, J.H., Patil, S.A.: Speech under stress: Analysis, modeling and recognition (2007)
15. Jabloun, F., Cetin, A.E., Erzin, E.: Teager energy based feature parameters for speech recognition in car noise. IEEE Sig. Process. Lett. **6**, 259–261 (1999)
16. Kumar, V., Minz, S.: Feature selection: a literature review. Smart CR **4**(3), 211–229 (2014)
17. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D., Choudhury, T.: Stresssense: detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp 2012, pp. 351–360. ACM, New York (2012). http://doi.acm.org/10.1145/2370216.2370270
18. Miller, T.C., Stone, D.N.: Public speaking apprehension (psa), motivation, and affect among accounting majors: a proofofconcept intervention. Issues Account. Educ. **24**(3), 265–298 (2009)
19. Sundaram, N., Smolenski, B., Yantorno, R.: Instantaneous nonlinear teager energy operator for robust voicedunvoiced speech classification (2003)
20. Pearson, R.K. (ed.): Exploring Data in Engineering, the Sciences, and Medicine. Oxford University Press, USA (2011)
21. Sarikaya, R., Gowdy, J.N.: Subband based classification of speech under stress. In: ICASSP, pp. 569–572 (1998)
22. Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bnziger, T.: Acoustic correlates of task load and stress. In: Hansen, J.H.L., Pellom, B.L. (eds.) INTERSPEECH. ISCA (2002)
23. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., Narayanan, S.: Paralinguistics in speech and language-state-of-the-art and the challenge. Comput. Speech Lang. **27**(1), 4–39 (2013)
24. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: INTERSPEECH, pp. 2253–2256. ISCA (2007)
25. Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: The interspeech 2012 speaker trait challenge. In: INTERSPEECH. ISCA (2012)
26. Sun, X.: A pitch determination algorithm based on subharmonic-to-harmonic ratio. In: the 6th International Conference of Spoken Language Processing, pp. 676–679 (2000)
27. Sun, Z., Li, Z.: Data intensive parallel feature selection method study. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 2256–2262, July 2014
28. Vogt, T., André, E., Wagner, J.: Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868, pp. 75–91. Springer, Heidelberg (2008)
29. Wells, J.: Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin (1997)
30. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Comput. **8**(7), 1341–1390 (1996)

31. Zhou, G., Hansen, J., Kaiser, J.: Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. **9**, 201–216 (2001)
32. Zuo, X., Fung, P.N.: A cross gender and cross lingual study on acoustic features for stress recognition in speech. In: Proceedings 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, pp. 2336–2339 (2011)