2016 ICSEE International Conference on the Science of Electrical Engineering

Detection of Distress in Speech

Yehav Alkaher, Osher Dahan, and Yair Moshe
Signal and Image Processing Laboratory (SIPL)
Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion – Israel Institute of Technology
Technion City, Haifa, Israel, http://sipl.technion.ac.il/

Abstract – **A distress situation is reflected in a person's speech e.g., fear, or in the speech of those around him e.g., anger. For many real-life situations, it may be desired to detect distress by remote monitoring. This paper deals with such remote monitoring using a microphone. Namely, we propose a technique for speaker-independent detection of distress in speech. Different temporal and spectral acoustic features such as the Mel frequency cepstral coefficients (MFCC) and the Teager energy operator (TEO) are investigated and the most relevant ones for the task are selected using the ReliefF feature selection algorithm. We use these features to train an SVM classifier in order to differentiate between speech in a distress situation and speech in which no distress is presented. On the Berlin Database of Emotional Speech, the proposed technique achieves classification accuracies of 91% per utterance and 87.8% per time window.**

Keywords* – *Distress detection; emotion recognition; emotional speech classification; speech analysis; Teager energy operator (TEO)

I. Introduction

Modern everyday life may introduce dangerous situations such as a robbery, an assault or a violent crime. In such situations, the victim feels in danger or suffers, and requires an immediate assistance. Most people nowadays carry a smartphone that can monitor their surrounding using a set of sensors. These sensors can be used to detect distress and automatically transmit an alarm signal to an emergency service or a nearby relative or friend. Each smartphone embeds a set of sensors that can be used together to detect distress. For example, a location (GPS) sensor to detect a dangerous neighborhood, accelerometers and a gyroscope to detect running or fall, a camera to detect an attack, or a microphone to detect explosions, gunshots, screams, threats or a cry for help. Efficient detection of distress situation is the goal of various systems. Most of them aim to monitor elderly people in their home indoor environment [1-3]. These systems use fusion between different sensors' data to detect a distress situation, such as falls, or sound events, such as glass breaking or screams [1, 2]. In these systems, distress in speech is usually detected by automatic speech recognition or by keyword spotting [3] to detect sentences of distress.

In this paper, we suggest a sound monitoring algorithm that detects distress acquired from a smartphone's microphone. Namely, we aim to detect fear or anxiety in a victim's voice or anger in an attacker's voice. We do not assume an indoor environment or a specific sensor. As in other systems, the result of the suggested technique may be fused with results of other audio processing algorithms or with analysis of data from other sensors on the device. A pre-processing stage is assumed to perform voice activity detection to separate speech from non-speech audio and to allow us to work only on speech signals. We do not analyze the signal to recognize words extracted from the speech or to extract informative keywords, thus only taking into account the emotion expressed in the speaker's voice.

Spoken text may have several interpretations, depending on how it is said. Therefore understanding the text alone is not sufficient to interpret the semantics of a spoken utterance [4]. The aim of emotion recognition in speech is to extract the non-linguistic emotional state of a speaker from his or her speech. This task is natural for humans but is a very challenging task for a machine [5]. This field of research is attracting a lot of attention recently. Emotion recognition can be viewed as a classification task into a discrete set of emotions. An important issue when aiming to solve this task is that no convention exists on defining the set of human emotions. Researchers have defined inventories of the emotional states encountered in our lives. Such an inventory may contains about 300 emotional states. However, classifying so many emotions is difficult. Many researchers agree with the 'palette theory', which states that any emotion can be decomposed into primary emotions similar to the way that any color is a combination of some basic colors. Primary emotions according to this theory are anger, disgust, fear, joy, sadness, and surprise [6]. Most previous works deal with emotion classification to one of these or similar primary emotions. In this paper, we deal with a binary distress classification, which is a special case of emotion classification in speech. We define distress by means of the emotions expressed in this kind of situations. Namely, we define distress as either anger or fear/anxiety.

The task of speech emotion recognition is very challenging for several reasons. First, it is not clear which speech features are most powerful in distinguishing between emotions [6]. The acoustic variability introduced by the existence of different utterances and speaking styles adds another obstacle because these properties affect most of the commonly extracted speech features such as pitch, and energy contours. Another challenging issue is that how a certain emotion is expressed depends on the speaker and language. Most works have tried to generalize by focusing on speaker-independent emotion classification, while other works are less generic but show

978-1-5090-2152-9/16/$31.00 ©2016 IEEE

better results by training a speaker-dependent classifier. Last, only few benchmark databases are publicly available and all these databases have limitations [6]. Thus, no standard speech corpora exist for comparing performance of different techniques used to recognize emotions [4].

As feature selection is a major issue in emotion recognition in speech, previous works investigate different acoustic features. Most researchers believe that prosody continuous features such as pitch and energy convey much of the emotional content of an utterance [6]. In addition, spectral features, such as the Mel-frequency cepstrum coefficients (MFCC), are often used as a short-time representation for speech signals. For detecting stress in speech, the Teager energy operator (TEO) was used in [7]. The selected features can be classified using different classifiers. The hidden Markov model (HMM) is the most used classifier in early works on emotion classification, probably due to its wide use for speech recognition applications. Other classifiers used for emotion classification are the Gaussian mixture model (GMM), artificial neural network (ANN), support vector machine (SVM), and k-nearest neighbors (kNN) [4-6].

We are not aware of any previous works that try to detect explicitly distress in speech. In [8], a set of temporal and spectral features are classified with an SVM classifier to achieves 95.1% emotion recognition accuracy on the Berlin emotional speech database (BESD). However, it classifies to three emotions (sadness, happiness and neutral), which are not associated with distress. Another work that uses a set of acoustic features with an SVM classifier is presented in [9]. This work reports on the BESD an accuracy of 82.5% for classification into five emotional states that cannot be explicitly associated with distress or no-distress. In [10], a set of 35 features was selected from a larger set of features by statistical methods and ANN, and random forest classifiers were used. Seven classes were categorized on the BESD. Analyzing the results of this paper in terms of distress detection indicates that it reaches 76% distress detection accuracy. GMM was used to classify emotions based on MFCC coefficients in [11]. Results are reported for six emotions on the BESD. When analyzed in terms of distress detection, the accuracy of that work is 80%. For stress recognition in speech, in [7] the performance of TEO-based features was evaluated compared with other acoustic features. The evaluation was performed on pronounced words. Stress was defined by angry, loud and Lombard speaking styles (Lombard reflex is the tendency of speakers to increase their vocal effort to increase communication quality when speaking in a noisy environment). It was shown that TEO-based features achieve a favorable detection accuracy of 97.9% on a dedicated database.

In this paper we propose a speaker-independent technique to detect the presence of distress (and not stress, in contradiction to [7]) in speech. Distress is defined as either anger or fear/anxiety. A set of acoustic linear and nonlinear features are extracted and we attempt to reduce the dimension of the feature space using principal component analysis (PCA) or using the ReliefF feature selection algorithm [12]. We train an SVM classifier and use it for binary classification between speech in which distress is presented and speech with no presence of distress. There are few novel aspects in this work. First, to the best of our knowledge, we are the first to deal explicitly with the problem of distress detection in speech. Another novelty is the use of both linear and non-linear (TEO-based) features. Last is the use of the ReliefF feature selection algorithm in the context of emotion detection.

The paper is organized as follows. In section II we describe the speech features under consideration, and in section III we describe the techniques used to select the most important features out of them. A general scheme of the suggested solution is given in section IV. We present our results in section V. Finally, in section VI we conclude our work.

II. Speech Features

One of the major issues in designing a speech emotion classifying algorithm is selecting the acoustic speech features to extract. This set of features should be efficient for emotion classification. Hence, on the one hand, it should contain information that characterizes the expressed emotion. On the other hand, it should be generic enough so it is not affected by variables such as speaker, gender or language. This section introduces the speech features used in this paper.

Speech features can be grouped into four categories [6]: continuous features, qualitative features, spectral features and TEO-based features. Many researchers believe that prosody continuous features such as pitch an energy disclose much of the emotional information of an utterance. Thus, using continuous features for emotion recognition has been sought in many studies. The features we extract in this category are the fundamental frequency (pitch) computed from the cepstrum, the root mean square (RMS) energy, the zero crossing rate, and the voice probability computed from the autocorrelation function.

Qualitative speech features are used to describe a speech signal by subjective voice quality labels, such as breathy or harsh. Several studies have been trying to define the relations between emotions and qualitative speech features. Although it is believed that these features have a strong correlation with the perceived emotion, they are difficult to extract, not reliable, and may differ with each researcher's interpretation. Therefore, relatively little is known about the role of voice quality in delivering emotions [6] and we decided not to include such features in our solution.

In addition to continuous acoustic features, various spectral features play a significant role in speech emotion recognition. These features inspect the short-time spectral envelope of a signal based on the understanding that the emotion content of an utterance can be exploited in the spectral domain. To approximate the human auditory system's response, the estimated spectral envelope of the signal is often passed through a bank of band-pass filters. Spectral features are then extracted from the outputs of these filters. The frequency scale can be linear as with the linear predictor cepstral coefficients (LPCC) or non-linear as with the

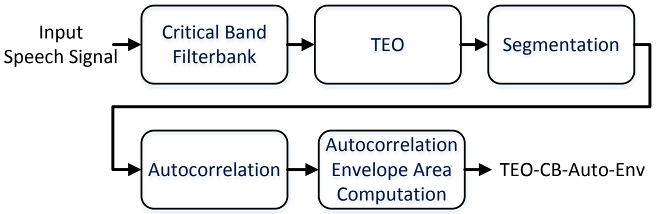

Fig. 1. Computation of the TEO-CB-Auto-Env feature [7] based on the Teager energy operator.

perceptual linear predictive (PLP) coefficients or the Mel frequency cepstral coefficients (MFCC). PLP uses the Bark-scale while MFCC uses the Mel-scale. Due to strong previous evidence regarding the usefulness of MFCC for speech emotion recognition, we extract 12 MFCC coefficients for spectral representation of the signal.

The Teager energy operator (TEO) [13, 14] can be used to detect stress in speech. TEO was derived according to a model that assumes speech is produced by a non-linear airflow in the vocal system. In different emotional conditions, the muscle tension of the speaker affects the airflow in the vocal system producing the sound. In addition, TEO was developed with the supporting evidence that hearing is the process of detecting energy [6]. The discrete time TEO is defined by

$$\Psi\left[x(n)\right]=x^{2}(n)-x(n+1)x(n-1) \qquad (1)$$

where $x(n)$ is a sample of the speech signal at time n . In [7], a set of TEO-based features are proposed for detecting neutral versus stressed speech. The authors concluded that a critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env) feature outperforms the pitch and the MFCC as well as the other proposed TEO-based features. The TEO-CB-Auto-Env is computed as depicted in Fig. 1. In order to correspond to the human auditory system, the signal is first filtered using a critical band based filterbank with 16 bands. Then, TEO is computed for each of the signal's frequency bands, and the output is segmented into time frames with a 50% overlap. Last, normalized TEO autocorrelation envelope area parameters are extracted for each critical band in each time frame. The rationale behind computing the autocorrelation envelope area is that this area is affected by pitch variation in the frame. If no pitch variation exists within a frame, the output TEO is a constant and its corresponding normalized autocorrelation function is a decaying straight line. When pitch variation is present in a frame, its normalized autocorrelation envelope is not an ideal straight line, and hence the area under the envelope will be smaller. By computing the area under the normalized autocorrelation envelope, we obtain 16 normalized TEO autocorrelation envelope area values for each time frame, one for each frequency band, which reflect the degree of excitation variability within each band [7]. Since TEO-CB-Auto-Env is reported to show good performance for detecting stress and since we assume partial correspondence between the expression of stress and distress in speech, we use TEO-CB-Auto-Env in this paper.

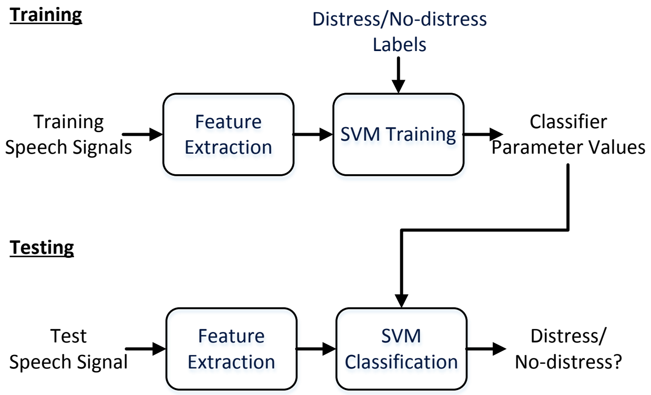

Fig. 2. The training and testing stages of the proposed technique for distress detection in speech.

III. Feature Selection

In machine learning, it is common to have redundancy in the feature space. In addition, some of the features may harm classification results by working as distracters or causing overfitting. To remove redundant or less important information, it is beneficial to reduce the number of random variables under consideration. This dimensionality reduction can be performed in one of two approaches - feature extraction or feature selection. We have investigated one dimensionality reduction algorithm from each approach – feature extraction by principal component analysis (PCA) [15] and feature selection by the ReliefF algorithm [12].

Feature extraction algorithms transform the data in the high-dimensional space to a space of fewer dimensions. PCA [15] is a common statistical procedure, in which a set of observations, with possibly correlated variables, is converted into a set of linearly uncorrelated principal components. Using a threshold operation on the variances of the principal components, less important components are eliminated.

Feature selection algorithms try to find a subset of the original variables without transforming the data. ReliefF [12] is a feature selection algorithm used in binary classification. It receives as input a dataset with n instances of p features, belonging to two known classes. The key idea of ReliefF is to estimate features importance according to how well their values distinguish among neighboring instances. An instance X is denoted by a p-dimensional vector $(x_1, x_2, \dots, x_p)$ where x_i denotes the value of feature f_i of X. Given x_i, ReliefF searches for its nearest two neighbors, one from the same class (called nearest hit) and the other from a different class (called nearest miss). The weight vector W_i of the instance x_i is then updated by:

$$W_i = W_i - \left|x_i - nearHit_i\right| + \left|x_i - nearMiss_i\right| \qquad (2)$$

Thus the weight of any feature decreases if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, and increases in the reverse case. After m iterations, the algorithm averages the contribution of nearest hits and misses to the weight vector to obtain the relevance vector. Features are selected if their relevance is greater than a threshold value.

IV. Distress Detection

Fig. 2. depicts a scheme of the proposed solution for distress detection in speech that follows the generic framework used for supervised classification. During training, feature extraction is used to convert each input speech signal to a feature set. Pairs of feature sets and labels are fed into a machine learning algorithm to generate a model, represented by the classifier parameter values. During testing, the same feature extraction is used to convert unseen speech signals to feature sets. These feature sets are then fed into the trained classifier (embedding the model), which generates predicted labels – “distress” or “no distress”. We use an SVM classifier since it was shown to provide good generalization performance in solving various classification problems.

We execute the proposed distress detection algorithm in two different configurations – global and local. In the global configuration, we classify speech utterances of variable length. In the local configuration, we divide each speech signal to small fixed-size time windows, each constitute of a set of frames, and classify each such window. A global configuration is superior to a local one in terms of classification accuracy. However, it cannot detect short-term expressions of distress and is not suitable for real-time implementation. As described in Section II, the set of acoustic features includes 32 scalars: pitch, energy, zero crossing rate, voice probability, 12 MFCC coefficients, and 16 TEO-CB-Auto-Env values. To capture the temporal information among neighboring time frames, we calculate for each feature the first order delta coefficient (derivative) by

$$d_t = \frac{\sum_{n=1}^{N} \left(f_{t+n} - f_{t-n} \right)}{2\sum_{n=1}^{N} n^2} \tag{3}$$

where d_t is the delta coefficient at time frame t and f_t is the feature value at time t. We use N=2. In total, the feature vector for each frame is of size 64.

In the global configuration, statistics of this feature vector over the utterance are calculated. Taking the statistics of the features over the frames captures important acoustic information while being robust to the variable length of utterances. Dozen statistics are calculated for each feature: minimum, maximum, range, position of minimum, position of maximum, arithmetic mean, the slope and the offset of a linear approximation of the contour, the quadric error of the linear regression, standard deviation, skewness and kurtosis. Thus, the size of the feature vector for each utterance is 64*12=768. In the local configuration, the feature vectors of all frames in each window are concatenated to compose a feature vector that represents the window. Since we use 10 frames per window, the size of the feature vector for each window is 32*10=640.

As explained in Section III, dimensionality reduction was performed on these sets of features. For the global and local configurations both PCA and and ReliefF where applied and the classification performance of the resulting features was compared. Results of this comparison are described in the Section V.

Table 1. Distress detection results in global configuration (per utterance) on the BESD for different combinations of features and dimensionality reduction techniques.

Linear Features	TEO-CB-Auto-Env	Dimension Reduction Technique	Vector Length [scalars]	Accuracy [%]
√	-	-	384	89.5
√	-	PCA	98	86.3
√	-	ReliefF	116	90.1
√	√	-	768	89.9
√	√	PCA	189	85.2
√	√	ReliefF	136	**91.0**

Table 2. Distress detection results in local configuration (per windows of 175 msec) on the BESD for different combinations of features and dimensionality reduction techniques.

Linear Features	TEO-CB-Auto-Env	Dimension Reduction Technique	Vector Length [scalars]	Accuracy [%]
√	-	-	320	87.3
√	-	PCA	290	80.1
√	-	ReliefF	160	87.2
√	√	-	640	86.9
√	√	PCA	120	74.7
√	√	ReliefF	440	**87.8**

V. Results

To evaluate the proposed distress detection technique, we have used the Berlin emotional speech database (BESD) [16]. This is a database in the German language that contains 10 utterances spoken by 10 actors in 7 primary emotions for a total of 535 utterances (not all utterances are spoken by all speakers in all emotions). Speech was recorded in an anechoic chamber having a good sound quality. This database was chosen since it was used in previous works, is compatible with our speaker-independent approach and is well documented. Tests were performed with frames of length 35 msec with an overlap of 50%. All features were calculated using the openSMILE library [17] except TEO-CB-Auto-Env which was implemented in MATLAB. For dimensionality reduction, either by PCA or by ReliefF, the threshold value for eliminating features was selected so high distress classification accuracy is achieved.

Table 1 lists results of the proposed distress detection technique in the global configuration namely classification per utterance. Performance was evaluated for six feature sets - with and without TEO-CB-Auto-Env, in three alternatives for dimensionality reduction - none, using PCA, and using ReliefF. To estimate the generalization performance of the proposed technique, a leave-one-out procedure was used. Accuracy is given by true detection rate. Highest accuracy of 91% is obtained by using both linear and TEO-based features with ReliefF. We can conclude from Table 1 that dimensionality reduction helps to improve results, and that ReliefF is superior in this aspect to PCA when using both linear and TEO-based features. As mentioned before, ReliefF results a relevance vector that measures the importance of each feature for classification. Note that in the feature set with highest accuracy, out of the 15 most important features according ReliefF, 12 are minimum or maximum of TEO-CB-Auto-Env value in a specific spectral band. In the feature set with only linear features and ReliefF, the most important feature is based on the signal energy and the second most important feature is based on the zero crossing rate. The following most important features in this feature set are statistics of MFCC coefficients.

Table 2 lists results of the proposed distress detection technique in the local configuration namely classification per window of fixed-size of 175 msec (10 frames of 35 msec each with overlap of 50%). Performance was evaluated in a similar way to the global configuration. Due to the large size of the training sets, we did not perform a leave-one-out procedure in this case but K-fold cross validation with K=4 (75% of the data used for training and 25% to testing in 4 iterations). Here again, highest accuracy is obtained by using both linear and TEO-based features with ReliefF. Accuracy in this case is 87.8%, lower than for the global configuration.

VI. Conclusions

This paper deals with speaker-independent distress detection in speech. A feature set is selected by considering a set of linear and non-linear (TEO-based) acoustic features and selecting the features that contribute the most to correct classification using the ReliefF feature selection algorithm. It was found that non-linear features are important and contribute to correct distress or no-distress classification. An SVM classifier is used in two configurations – global (classification per utterance) and local (classification per fixed-size window). On the Berlin Database of Emotional Speech, classification accuracies of 91% were obtained in the global configuration and 87.8% in the local configuration.

Acknowledgment

The authors would like to thank Prof. David Malah, head of SIPL, and Nimrod Peleg, chief engineer of SIPL, for their support, advice and helpful comments. The authors would also like to thank Dr. Gideon Meiry from cMeSafe for the collaboration and support.

References

[1] C. Doukas, L. Athanasiou, K. Fakos, and I. Maglogiannis, "Advanced sound and distress speech expression classification for human status awareness in assistive environments," *The Journal on Information Technology in Healthcare,* vol. 7, pp. 111-117, 2009.

[2] J. Rougui, D. Istrate, and W. Souidene, "Audio sound event identification for distress situations and context awareness," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 3501-3504.

[3] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Speech recognition of aged voice in the aal context: Detection of distress sentences," in *Speech Technology and Human-Computer Dialogue (SpeD), 2013 7th Conference on*, 2013, pp. 1-8.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology,* vol. 15, pp. 99-117, 2012.

[5] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 2, pp. 235-238, 2012.

[6] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, pp. 572-587, 2011.

[7] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing,* vol. 9, pp. 201-216, 2001.

[8] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home,* vol. 6, pp. 101-108, 2012.

[9] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*, 2011, pp. 621-625.

[10] T. Iliou and C.-N. Anagnostopoulos, "Comparison of different classifiers for emotion recognition," in *Informatics, 2009. PCI'09. 13th Panhellenic Conference on*, 2009, pp. 102-106.

[11] M. M. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 2007, pp. IV-957-IV-960.

[12] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence,* vol. 7, pp. 39-55, 1997.

[13] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, pp. 599-601, 1980.

[14] L. Kaiser, "Communication of affects by single vowels," *Synthese,* vol. 14, pp. 300-319, 1962.

[15] C. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn," ed: Springer, New York, 2007.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, 2005, pp. 1517-1520.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462.