THE EFFECT OF EMOTIONAL SPEECH ON RELATIVE SPEAKER DISCRIMINATION

Juliane Höbel-Müller, Ronald Böck, Andreas Wendemuth

Institute for Information and Communications Engineering, Cognitive Systems Group, Otto-von-Guericke University Magdeburg
juliane.hoebel@ovgu.de

Abstract: Text-independent speaker discrimination (SD) involves checking whether two arbitrary speech signals are uttered by the same speaker or two different speakers. It has various applications such as speaker verification or speech turn segmentation. However, emotionally colored speech introduces variations in the acoustic features impairing the performance of baseline speech technologies. This study focuses on investigating the influence of emotions on SD, applying an approach based on a relative characterization of the speaker, called Relative Speaker Characteristic (RSC). The intrinsic variability is modeled by using emotional utterances represented in the benchmark corpus Berlin Database of Emotional Speech. Three feature subsets based on Mel Frequency Cepstral Coefficients (MFCCs) are used to calculate the RSC that represents the SD specific information, namely $\mathscr{F}_1 = \{13 \text{ MFCCs}\}$, $\mathscr{F}_2 = \{\mathscr{F}_1 \cup \text{ delta coefficients}\}$ and $\mathscr{F}_3 = \{\mathscr{F}_2 \cup \text{ delta} - \text{delta coefficients}\}$. Emotionally neutral utterances serve as training data. SD models are developed using a Support Vector Machine with a linear kernel. By using the RSC that is based on $\mathscr{F}_1$, the best SD performance is achieved. Regarding $\mathscr{F}_1$, the SD performance for utterances in the state of joy (EER= 6.6%), boredom (EER= 6.69%) and anger (EER= 7.61%) is similar to the SD for emotionally neutral utterances (EER= 7.34%). However, for utterances in the state of fear (EER = 10.91%), disgust (EER= 23.76%) and sadness (EER= 25.76%), the SD performance is unreliable.

1 Introduction

Text-independent acoustic speaker discrimination (SD) is termed as checking whether two arbitrary speech signals are uttered by the same speaker or by two different speakers [1]. There are various speech technologies based on SD, such as speaker identification, speaker verification [2] and speech turn segmentation [3]. Furthermore, there is a growing need to apply automatic SD to speech signals to provide information that could be used in assistance systems [4]. Since in many systems the speech is commonly emotional [5], several studies prove that emotional speech strongly degrades the performance of current speech technologies [6, 7]. These studies used acted emotional speech. As shown in [8] for speaker verification relying on the MSP-PODCAST corpus, performance also declines for naturalistic utterances with extreme values for arousal and valence representing emotional attributes. Multi-modal assistance systems are being developed for SD; for instance, using microphone arrays and cameras for emotion-independent audio-visual speaker recognition [9]. However, this kind of SD remains vague due to limited audio source location accuracy. In order to enhance the adaptability at the algorithmic level, a robust text-independent acoustic speaker modeling technique should be applied.

Previous studies - mostly using acted emotional speech - show that the speaker modeling techniques' performances degrade when the models are trained on emotionally neutral speech and tested on emotional speech uttered happy, angry or sad [10, 11, 7]. For instance, Gaussian mixture modeling (GMM) techniques and Support Vector Machines (SVM) using Mel Frequency Cepstral Coefficients (MFCCs) are strongly degraded by emotional speech [11, 12]. Hidden Markov Models (HMMs), Second-Order Circular Hidden Markov Models (CHMM2s) and Suprasegmental Hidden Markov Models (SPHMMs) with Log-Frequency Power Coefficients (LFPC) have been tested by Shahin [13], who shows that SPHMMs perform best compared to other models. Furthermore, Shahin proposes a two-stage recognizer comprising an HMM-based emotion classifier and an SPHMM-based speaker classifier for speaker identification in affective environments [14]. The accuracy based on the proposed two-stage recognizer is significantly improved compared to a one-stage recognizer [14]. Moreover, by applying a fusion of GMM-based classifiers with MFCCs, Line Spectral Frequencies and Temporal Energy of Subband Cepstral Coefficients, speaker identification can be enhanced in both neutral and emotional speech [15].

Another approach for developing emotion-independent speaker identification is proposed by Wu et al. [6], introducing an emotion-dependent score normalization that copes with GMM-universal background model (UBM) based score variability. Furthermore, emotion-added models [16] and emotion-state conversion [17] have been applied for affect compensation. Bao et al. [10] proposed an affect compensation method called emotion attribute projection by adapting a channel compensation method. Moreover, feedforward neural networks [18] and auto-associative neural networks [19] have been applied for feature transformation. When emotional speech's features are transformed into their emotionally neutral equivalents, speaker identification accuracy on acted emotional speech can be significantly improved compared to approaches using no transformation [19].

Ouamour et al. [1] proposed a feature reduction technique based on a relativistic approach. The authors investigated a relative characterization of the speaker, called Relative Speaker Characteristic (RSC), which is suitable for SD in a noisy environment or for telephonic speech [1]. This study focus on investigating the influence of emotions on SD, applying the RSC. Accordingly, the following research question arises: Is the RSC-based SD's [1] performance influenced by emotional speech in comparison with emotionally neutral speech? To answer this question, we present a performance evaluation of a text-independent RSC-based SD system, first, trained and tested on emotionally neutral utterances, and second, tested on emotional speech.

2 Corpus and Feature Extraction

2.1 Emotional Speech Corpus

The benchmark speech database used in this study is the Berlin Database of Emotional Speech (EMO-DB) [20]. The speakers are ten actors (five females). The female actors are on average 30.6 ± 5.6 years old and the male actors are on average 28.8 ± 3.1 years old. Each of them simulates different emotional states when asked to recite ten different German utterances with neutral semantic content. Overall, the database contains 494 different utterances labeled with the following seven emotional states: anger, boredom, disgust, fear, joy, neutral and sadness. The recordings sampled at 16 kHz provide a high audio quality, minimizing extrinsic variability factors.

2.2 Feature Extraction

For each EMO-DB utterance, three feature subsets were calculated, namely $\mathscr{F}_1 = \{13\,\text{MFCCs}\}$, $\mathscr{F}_2 = \{\mathscr{F}_1 \cup \text{delta coefficients}\}$ and $\mathscr{F}_3 = \{\mathscr{F}_2 \cup \text{delta} - \text{delta coefficients}\}$, provided by openSMILE [21], where a feature vector was computed with 20 ms frame length every 10 ms. Normalization was performed for each utterance using mean and standard deviation. The feature vectors based on either $\mathscr{F}_1$, $\mathscr{F}_2$ or $\mathscr{F}_3$ were used to calculate covariance matrices. By using the covariance matrices, $\text{RSC}_{\mathscr{F}_1}$ based on $\mathscr{F}_1$, $\text{RSC}_{\mathscr{F}_2}$ based on $\mathscr{F}_2$, and $\text{RSC}_{\mathscr{F}_3}$ based on $\mathscr{F}_3$ are calculated. Only $\text{RSC}_{\mathscr{F}_i}$ with $i \in \{1,2,3\}$ is used in the classification experiments and modeling, respectively.

Given a set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ of feature vectors resulting from the acoustic analysis of a speech signal u_x, the corresponding covariance matrix $\mathbf{X}$ can be calculated. Similarly, the covariance matrix $\mathbf{Y}$ is extracted for a set of feature vectors $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m\}$ resulting from the acoustic analysis of a speech signal u_y. By considering information provided by covariance matrices, RSC is calculated constructing a meta-feature space. The RSC represents a similarity measure incorporating the relative statistics of a speaker compared with another speaker, which is considered as a reference speaker [1]. According to [1], the RSC is defined as a matrix

$$\text{RSC}(u_x, u_y) = \mathbf{X}^{-1} \cdot \mathbf{Y}, \tag{1}$$

where $\mathbf{X}$, $\mathbf{Y}$ denotes the covariance matrices of u_x, u_y.

In order to enhance the SD information, Ouamour et al. [1] argued for combining $\text{RSC}(s_x, s_y)$ and $\text{RSC}(s_y, s_x)$ (see Equation 2).

$$\text{RSC}_{(}u_x, u_y) = [\mathbf{X}^{-1} \cdot \mathbf{Y} \;\; \mathbf{Y}^{-1} \cdot \mathbf{X}] \tag{2}$$

Moreover, Ouamour et al. [1] hypothesize that the most important information for SD is located in the principal diagonal of $\text{RSC}(u_x, u_y)$ in Equation 1. The authors stated the following three cases: first, RSC will be the identity matrix if the two utterances are identical and uttered by the same speaker; second, RSC's distance to the identity matrix will be relatively small if the different utterances belong to the same speaker; and third, RSC will generate large values in the matrix' diagonal compared with the non-diagonal elements' values if the utterances are obtained by two different speakers. Hence, calculating the RSC is modified in Equation 3 [1].

$$\text{RSC}(u_x, u_y) = [\text{diag}(\mathbf{X}^{-1} \cdot \mathbf{Y}) \;\; \text{diag}(\mathbf{Y}^{-1} \cdot \mathbf{X})] \tag{3}$$

Since the most important information for SD is located in the diagonal [1], we simplify the expression defined in Equation 3 by using only the variance of each variable (see Equation 4).

$$\text{RSC}(u_x, u_y) = [\text{diag}(\mathbf{Var}_{u_x}^{-1} \cdot \mathbf{Var}_{u_y}) \;\; \text{diag}(\mathbf{Var}_{u_y}^{-1} \cdot \mathbf{Var}_{u_x})], \tag{4}$$

where $\mathbf{Var}_{u_x}$ is the variance matrix for u_x defined in Equation 5.

$$\mathbf{Var}_{u_x} = \begin{bmatrix} \sigma_1^2(u_x) & 0 & \dots & 0 \\ 0 & \sigma_2^2(u_x) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_D^2(u_x) \end{bmatrix} \tag{5}$$

Accordingly, the variance matrix $\mathbf{Var}_{u_y}$ for u_y is calculated. The variances ${\sigma_i}^2(u_x)$ are extracted from the variance vector given as

$$[\sigma_1^2(u_x) \;\; \sigma_2^2(u_x) \dots \sigma_D^2(u_x)] = \frac{1}{n-1} \sum_{i=1}^{n} \mid \mathbf{x}_i - \overline{\mathbf{x}} \mid^2, \tag{6}$$

where $\bar{\mathbf{x}}$ denotes the mean feature vector and D denotes the feature vector's dimension. Similarly, the variances for an utterance u_y are extracted.

Preprocessing u_x such that there are no pauses in the utterance is a prerequisite to calculate the RSC based on either $\mathscr{F}_1$, $\mathscr{F}_2$ or $\mathscr{F}_3$. Subsequently, the *invertible* matrices $\mathbf{Var}_{u_x}$ and $\mathbf{Var}_{u_y}$ are matrices such that $\sigma_i^2(u_x) \neq 0$ for all preprocessed u_x. As $\mathbf{Var}_{u_x}$ is a diagonal matrix with non-zero principal diagonal elements, a simplification is undertaken by only calculating the inverse elements of $\mathbf{Var}_{u_x}$'s principal diagonal, which is equivalent to calculating $\mathbf{Var}_{u_x}^{-1}$. In summary, feature extraction starts with calculating $\mathscr{F}_1$, $\mathscr{F}_2$ and $\mathscr{F}_3$, and generates $\text{RSC}_{\mathscr{F}_1}$, $\text{RSC}_{\mathscr{F}_2}$ and $\text{RSC}_{\mathscr{F}_3}$. Only $\text{RSC}_{\mathscr{F}_i}$ with $i \in \{1,2,3\}$ is used in the classification experiments and modeling, respectively. In the last step of the feature extraction process, standardization was utilized as the RSC is a non-standardized measure.

3 Experimental Design

The two SD classes *different speaker* and *same speaker* are modeled with a SVM with a linear kernel using the RSC (see Equation 4). An SVM is chosen since previous studies show that GMMs do not provide suitable results for SD tasks [22]. The SVM classification score for classifying an RSC feature vector is computed using Matlab.

Three experiments are conducted using $\text{RSC}_{\mathscr{F}_1}$, $\text{RSC}_{\mathscr{F}_2}$ and $\text{RSC}_{\mathscr{F}_3}$ introduced in section 2. Emotionally neutral utterances from EMO-DB served as training data in the experiments. A ten-fold cross-validation was conducted considering only emotionally neutral utterances to obtain a baseline. By evaluating the influence of each emotion on SD, we answered our research question. Thereunto, both detection error trade-off (DET), which gives the minimum cost, according to a Detection Cost Function, and DET curves are applied using NIST's DET-Curve Plotting software[1]. The DET giving the minimum cost corresponds to the equal error rate (EER).

4 Speaker Discrimination Results on Emotional Speech and Discussion

In this section, three experiments concerning the influence of emotion on the RSC-based SD are presented and discussed.

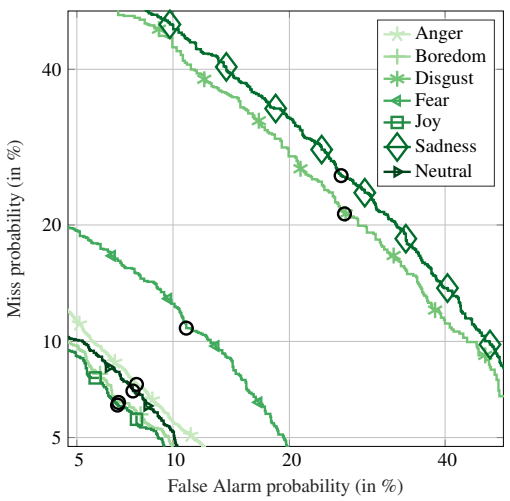
Including emotionally neutral speech, the average EERs for all emotions regarding each feature set are

- $\overline{\text{EER}}_{\text{RSC}_{\mathscr{F}1}} = 12.67 \pm 8.41\%$,
- $\overline{\text{EER}}_{\text{RSC}_{\mathscr{F}2}} = 13.83 \pm 8.01\%$ and
- $\overline{\text{EER}}_{\text{RSC}_{\mathscr{F}3}} = 31.97 \pm 12.88\%$.

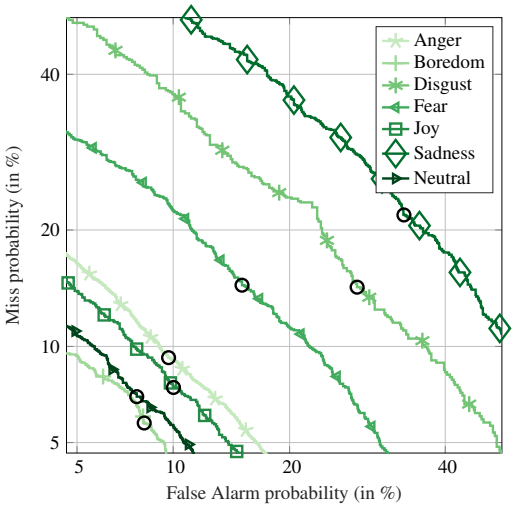
In accordance with [11, 12], we can state that $\text{RSC}_{\mathscr{F}1}$ and $\text{RSC}_{\mathscr{F}2}$ are more descriptive than $\text{RSC}_{\mathscr{F}3}$ in SD considering a small gender-balanced set of ten speakers. However, it can be hypothesized that the discriminative power of $\text{RSC}_{\mathscr{F}3}$ increases in a dataset with a higher number of speakers. This assumption is in accordance with [8] considering the MSP-PODCAST corpus with 40 speakers, using $\mathscr{F}_3$ in an i-vector framework and reporting mean EERs under 5% for naturalistic speech in the state of emotions with extreme values for the arousal and valence scores.

According to Figures 1a and 1b, the EERs can be clustered into two categories. The first category comprises similar EERs arising from speech in the emotions of neutral, boredom, joy

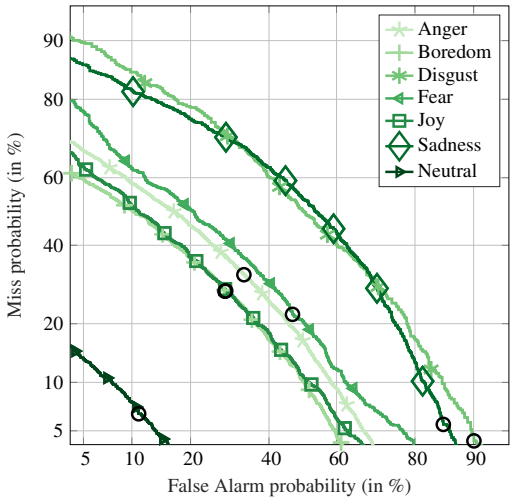
[1] https://www.nist.gov/itl/iad/mig/tools (accessed January 22, 2018)



(a) Results of SVMs trained on $\mathrm{RSC}_{\mathscr{F}_1}$.



(b) Results of SVMs trained on $\mathrm{RSC}_{\mathscr{F}_2}$.



(c) Results of SVMs trained on $\mathrm{RSC}_{\mathscr{F}_3}$.

Figure 1 – DET curves presenting results of SVMs trained on (a) $\mathrm{RSC}_{\mathscr{F}_1}$, (b) $\mathrm{RSC}_{\mathscr{F}_2}$, and (c) $\mathrm{RSC}_{\mathscr{F}_3}$ extracted from emotionally neutral utterances, and tested on utterances in the emotions of anger, boredom, disgust, fear, joy, neutral and sadness. The black circles show the minimum cost points and correspond to the EER.

and anger, where neutral can be understood as centroid of this EER cluster. EERs arising from speech in the emotions of fear, disgust and sadness can be considered as outliers. Regarding Figures 1a and 1b, the $\mathrm{EERs}_{\mathrm{RSC}_{\mathscr{F}1}}$ are smaller and closer together compared to the $\mathrm{EERs}_{\mathrm{RSC}_{\mathscr{F}2}}$. As shown in Figure 1c, classification which is based on $\mathrm{RSC}_{\mathscr{F}3}$ results in unreliable EERs arising from speech in anger, boredom, disgust, fear, joy and sadness. The $\mathrm{EER}_{\mathscr{F}3}$ arising from emotionally neutral speech appears only as an outlier in Figure 1c. This fact makes the $\mathrm{RSC}_{\mathscr{F}3}$ for SD unsuitable.

Considering the best suitable feature set $\mathscr{F}_1$ the $\mathrm{EERs}_{\mathrm{RSC}_{\mathscr{F}1}}$ for speech in the emotions of fear (10.91%), disgust (23.76%) and sadness (25.76%) increase more compared to speech in the emotions of joy (6.6%), boredom (6.69%) and anger (7.61%). A reason is given in a high value for arousal in the emotions of fear and disgust [23]. Moreover, sadness has a high negative valence in comparison with neutral. In the state of the art, there are no studies analyzing the effect of emotion on SD. Therefore, we consider study results in the following regarding baseline speaker verification. Wu et al. [6] and Ghiurcau et al. [11] analyzed the influence of emotions on a GMM-UBM system using MFCCs and delta coefficients calculated on an acted emotional speech corpus. Compared to the EER in a test of emotionally neutral utterances (4.48%) in

[6], Wu et al. [6] reported that the EERs increased when utterances in the emotions of sadness (12.59%), joy (17.24%), anger (17.93%) and fear (18.62%) were tested. In contrast to [6], Ghiurcau et al. [11] presented only the accuracy for speech in the emotions of sadness (90%), boredom (90%), fear (52%), joy (36%) and anger (36%), respectively. Ghiurcau et al. [12] repeated their previous study described in [11] using an SVM with polynomial kernel rather than a GMM-UBM system. The authors obtained similar results compared to their previous study in [11]. It is debatable to compare the results in [6] with those stated in [11] (or in [12]) as the authors use different performance measures (EER [6] and accuracy [11, 12]), and their analysis is based on different corpora (EMO-DB [11, 12] and an acted Chinese speech corpus [6]). However, both works confirmed that speech in the emotions of anger highly degrades the performance of speaker verification compared to emotionally neutral speech. In our work, speech in the emotions of fear, disgust and sadness represents the most difficult emotional speech for SD, whereas the SD is only marginally degraded by speech in the emotions of anger, boredom and joy. Therefore, the $\text{RSC}_{\mathscr{F}i}$ with $i \in \{1,2,3\}$ is not a suitable speaker-specific measure when speech is in the emotions of fear, disgust or sadness.

5 Conclusion and Outlook

In this work, the effect of emotions on RSC-based SD has been evaluated. It has been proved that emotion involved in the testing utterances will aggravate the SD performance. In contrast to baseline approaches, the RSC-based SD shows small performance loss when utterances affected by boredom, joy or anger were tested. However, speech uttered in fear, disgust or sadness highly degrades the SD performance, which is partly in accordance with tests of baseline systems (cf. [6, 11, 12]). A reason for this performance decline is the large distance between those emotions and neutral in the valence-arousal space. Moreover, the classification performance is affected by the small number of utterances in those emotions compared to the number of utterances in the emotions of joy, anger or boredom.

In comparison with GMM-based systems, the RSC-specific mapping was captured from the training data itself, and no assumption about the underlying probability density functions is given. Another advantage of the proposed method is that the number of free parameters is small. Using $\text{RSC}_{\mathscr{F}1}$, $\text{RSC}_{\mathscr{F}2}$ or $\text{RSC}_{\mathscr{F}3}$, the number of free parameters in the RSC approach is 23, 52, or 78, respectively. Given the simplifications in Equation 4, the simplified RSC approach is computationally more efficient than the original RSC approach presented in [1].

Multi-modal assistance systems should robustly handle emotions with a high negative value for valence such as sadness, or with a high positive value for arousal such as fear and disgust, to adequately react to users' distress situations. However, RSC-based SD's reliability is degraded by speech in those emotions. Therefore, as proposed in [14] for standard speaker identification, applying a two-stage recognizer comprising an emotion classifier and a speaker classifier could be suitable. Additionally, Ouamour et al. [1] showed that SD is degraded by models that are based on mixed genders in contrast to mono-gender training and testing. Therefore, a gender-dependent training and a gender recognizer can enhance the SD [1]. In order to simulate a realistic scenario, a cross-corpus analysis will be applied using naturalistic emotional speech.

6 Acknowledgement

We acknowledge support by the project "Intention-based Anticipatory Interactive Systems" (IAIS) funded by the Federal State of Sachsen-Anhalt, Germany. Further, we thank the project "Mova3D" (grant number: 03ZZ0431H) funded by 3Dsensation within the Zwanzig20 funding program by the German Federal Ministry of Education and Research (BMBF).

References

[1] OUAMOUR, S., M. GUERTI, and H. SAYOUD: *A new relativistic vision in speaker discrimination*. *Canadian Acoustics Journal*, 36(4), pp. 24–35, 2008.

[2] BULGAKOVA, E., A. SHOLOHOV, N. TOMASHENKO, and Y. MATVEEV: *Speaker verification using spectral and durational segmental characteristics*. In A. RONZHIN, R. POTAPOVA, and N. FAKOTAKIS (eds.), *Speech and Computer*, pp. 397–404. Springer International Publishing, Athens, Greece, 2015.

[3] INDIA, M., J. A. FONOLLOSA, and J. HERNANDO: *Lstm neural network-based speaker segmentation using acoustic and language modelling*. In *Proc. of the Interspeech-2017*, pp. 2834–2838. Stockholm, Sweden, 2017.

[4] VACHER, M., F. PORTET, A. FLEURY, and N. NOURY: *Challenges in the processing of audio channels for ambient assisted living*. In *The 12th IEEE International Conference on e-Health Networking, Applications and Services*, pp. 330–337. 2010.

[5] PICARD, R. W.: *Affective computing: challenges*. *International Journal of Human-Computer Studies*, 59(1), pp. 55–64, 2003.

[6] WU, W., T. F. ZHENG, M.-X. XU, and H.-J. BAO: *Study on speaker verification on emotional speech*. In *Proc. of the Interspeech-2006*. 2006.

[7] SHAHIN, I.: *Speaker identification in emotional talking environments based on csphmm2s*. *Engineering Applications of Artificial Intelligence*, 26(7), pp. 1652 – 1659, 2013.

[8] PARTHASARATHY, S., C. ZHANG, J. H. L. HANSEN, and C. BUSSO: *A study of speaker verification performance with expressive speech*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5540–5544. 2017.

[9] MAGANTI, H. K., D. GATICA-PEREZ, and I. MCCOWAN: *Speech enhancement and recognition in meetings with an audio-visual sensor array*. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8), pp. 2257–2269, 2007.

[10] BAO, H., M.-X. XU, and T. F. ZHENG: *Emotion attribute projection for speaker recognition on emotional speech*. In *Proc. of the Interspeech-2007*. Antwerp, Belgium, 2007.

[11] GHIURCAU, M. V., C. RUSU, and J. ASTOLA: *A study of the effect of emotional state upon text-independent speaker identification*. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4944–4947. IEEE, Prague, Czech Republic, 2011.

[12] GHIURCAU, M. V., C. RUSU, and J. ASTOLA: *Speaker recognition in an emotional environment*. In *Proc. of Signal Processing and Applied Mathematics for Electronics and Communications (SPAMEC)*, pp. 81–84. Cluj-Napoca, Romania, 2011.

[13] SHAHIN, I.: *Speaker identification in emotional environments*. *Iranian Journal of Electrical and Computer Engineering*, 8(1), pp. 41–46, 2009.

[14] SHAHIN, I.: *Identifying speakers using their emotion cues*. *International Journal of Speech Technology*, 14(2), pp. 89–98, 2011.

[15] JAWARKAR, N. P., R. S. HOLAMBE, and T. K. BASU: *Frontiers in Computer Education*, chap. Text-Independent Speaker Identification in Emotional Environments: A Classifier Fusion Approach, pp. 569–576. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[16] LI, D., Y. YANG, Z. WU, and T. WU: *Emotion-state conversion for speaker recognition*. In J. TAO, T. TAN, and R. W. PICARD (eds.), *Affective Computing and Intelligent Interaction*, pp. 403–410. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[17] WU, T., Y. YANG, and Z. WU: *Affective Computing and Intelligent Interaction (ACII 2005)*, chap. Improving Speaker Recognition by Training on Emotion-Added Models, pp. 382–389. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[18] KROTHAPALLI, S. R., J. YADAV, S. SARKAR, S. G. KOOLAGUDI, and A. K. VUPPALA: *Neural network based feature transformation for emotion independent speaker identification*. *International Journal of Speech Technology*, 15(3), pp. 335–349, 2012.

[19] KOOLAGUDI, S. G., S. E. FATIMA, and K. S. RAO: *Speaker recognition in the case of emotional environment using transformation of speech features*. In *Proc. of the CUBE International Information Technology Conference*, CUBE '12, pp. 118–123. ACM, New York, NY, USA, 2012.

[20] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A database of german emotional speech*. In *Proc. of the Interspeech-2005*, pp. 1517–1520. Lissabon, Portugal, 2005.

[21] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE – the munich versatile and fast open-source audio feature extractor*. In *Proc. of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462. ACM, New York, NY, USA, 2010.

[22] SAYOUD, H., S. OUAMOUR, and Z. HAMADACHE: *Discriminating speakers by their voices — a fusion based approach*. In A. KARPOV, R. POTAPOVA, and I. MPORAS (eds.), *Speech and Computer-2017*, pp. 322–331. Springer International Publishing, Hatfield, UK, 2017.

[23] COWIE, R., E. DOUGLAS-COWIE, N. TSAPATSOULIS, G. VOTSIS, S. KOLLIAS, W. FELLENZ, and J. G. TAYLOR: *Emotion recognition in human-computer interaction*. *IEEE Signal Processing Magazine*, 18(1), pp. 32–80, 2001.