A STUDY OF SPEAKER VERIFICATION PERFORMANCE WITH EXPRESSIVE SPEECH

Srinivas Parthasarathy, Chunlei Zhang, John H.L. Hansen and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory - Center for Robust Speech Systems (CRSS)
Department of Electrical Engineering, The University of Texas at Dallas, Richardson TX 75080, USA
sxp120931@utdallas.edu, cxz142930@utdallas.edu, john.hansen@utdallas.edu, busso@utdallas.edu

ABSTRACT

Expressive speech introduces variations in the acoustic features affecting the performance of speech technology such as speaker verification systems. It is important to identify the range of emotions for which we can reliably estimate speaker verification tasks. This paper studies the performance of a speaker verification system as a function of emotions. Instead of categorical classes such as happiness or anger, which have important intra-class variability, we use the continuous attributes arousal, valence, and dominance which facilitate the analysis. We evaluate an speaker verification system trained with the i-vector framework with a *probabilistic linear discriminant analysis* (PLDA) back-end. The study relies on a subset of the MSP-PODCAST corpus, which has naturalistic recordings from 40 speakers. We train the system with neutral speech, creating mismatches on the testing set. The results show that speaker verification errors increase when the values of the emotional attributes increase. For neutral/moderate values of arousal, valence and dominance, the speaker verification performance are reliable. These results are also observed when we artificially force the sentences to have the same duration.

Index Terms— Speaker verification, emotion recognition

1. INTRODUCTIONS

Expressive speech introduces challenges in current speech technologies. The externalization of emotions produces deviations from neutral speech reflected on the glottal waveform, prosody, spectral characteristics, and speech duration [1–3]. As a result, the performance of *automatic speech recognition* (ASR) [4, 5] and speaker recognition/verification [6–9] significantly degrades with expressive speech. For speaker verification systems, the drop in performance is a problem, since in many applications the speech of interest is commonly emotional.

Several studies have reported a drop in speaker verification identification performance in mismatched conditions when the models are trained with neutral speech and tested with emotional speech [7, 9]. Previous studies have proposed compensation schemes to attenuate this problem [7, 8, 10–14]. An important limitation of previous speaker verification studies on expressive speech is the database used for the analysis. The studies have relied on acted databases where speakers were asked to repeat utterances conveying different emotions [6, 9, 15, 16]. Furthermore, the emotional databases used for this purpose are recorded from limited number of speakers (e.g., less than 10) [9, 15, 16], which are not enough to reliably study the performance of speaker verification frameworks. It is important to understand the effect of emotion in speaker verification tasks under naturalistic conditions when the number of speaker is large.

This work was funded by NSF CAREER award IIS-1453781.

This paper analyzes the effect of emotion on speaker verification tasks. It uses a portion of the MSP-PODCAST corpus, which is a large emotional speech database that is being created at the University of Texas at Dallas. The corpus contains many hours of recordings from several speakers appearing on creative common licensed podcasts. The emotional content of the corpus is annotated with crowdsourcing evaluations. The analysis relies on a subset of 40 speakers with varied emotional content. Using a state-of-the-art system, we analyze speaker verification performance in terms of the emotional attributes arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong). The results show an increase of speaker verification error rate as the arousal, valence, dominance levels of the sentences increase, especially for extreme emotional values. The error rates strongly depend on the duration of the sentence, where the error rate increases for shorter sentences (e.g., less than 4s). To compensate for decrease in performance due to duration, we repeated the analysis for sentences with same duration (5s). The results identify the areas in the arousal-valence space where the speaker verification performance is more affected by emotional deviations. The MSP-PODCAST corpus and the analysis in this study opens interesting research questions on affective computing in the area of speaker verification.

2. RELATION TO PRIOR WORK

Previous studies have analyzed the effect of different emotional categories such as happiness, anger, and sadness on speaker verification systems, proposing compensation schemes to increase the performance with expressive speech [7, 10]. Li et al. [8] and Wu et al. [11] proposed features modifications from neutral speech to different emotional categories to improve performance of speaker identification systems. Krothapalli et al. [12] used neutral networks to transform features from emotional categories to neutral domain before training speaker identification systems. Shahin [13] used both emotional and gender cues to train speaker identification systems. Li and Yang [14] proposed an alternative approach, where they match the emotional state between test and train utterances by clustering affective speech. Then, they built corresponding models to improve performance on speaker recognition.

Previous studies used emotional corpora with either acted recordings or from limited number of speakers. There is a need for an emotional database that is suitable to systematically analysis speaker verification performance in the presence of emotion. The next section describes the MSP-PODCAST, which satisfies this requirement.

3. DATABASE

The study relies on the MSP-PODCAST dataset currently being created at The University of Texas in Dallas. This speech corpus is part

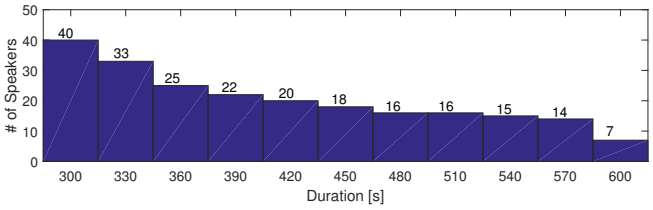


Fig. 1. Number of speakers with at least d seconds in the dataset. For example, 16 speakers have more than eight minutes (i.e., 480s).

of a National Science Foundation project aiming to develop emotion recognition systems that generalize across conditions. The dataset comprises multiple sentences from different speakers appearing in various podcasts that are publicly available under Creative Commons license. The durations of the selected sentences range from 2.75s to 11s. A big advantage of this corpus is the availability of large number of naturalistic, emotionally colored sentences, which are annotated by at least five raters on Amazon Mechanical Turk using a variation of the online assessment framework proposed in Burmania et al. [17]. The perceptual evaluation includes seven-point Likert-scales for the emotional attributes arousal (1- very calm versus 7 - very active), valence (1- very negative versus 7- very positive) and dominance (1- very weak versus 7- very strong). The evaluation also includes primary categorical emotions where the raters selected the most appropriate class (happiness, sadness, anger, fear, contempt, disgust, surprised, neutrality and other). The consensus label is obtained with majority vote rule. Currently, the corpus has 7,070 sentences (approximately 11 hours).

For the speaker verification experiments, the study considers a subset of this corpus. We manually identified recordings from 40 speakers, where each of them has over 300s of speech recordings. Figure 1 gives the number of speakers with at least d seconds. For example, there are 18 subjects with more than 450s. The selected set has 3025 sentences (4hrs 57min).

4. SPEAKER VERIFICATION FRAMEWORK

In this study, we employ an i-vector framework with a *probabilistic linear discriminant analysis* (PLDA) back-end to test how emotional dimensional attributes influence speaker verification performance [18, 19].

4.1. I-vector modeling

In speaker recognition research, the i-vector scheme provides an elegant way of reducing a *maximum a posteriori* (MAP) adapted high-dimensional Gaussian supervector into a low-dimensional vector, while retaining most of the high-level information of a speech segment. This framework provides a suitable front-end due to the rich information and the fixed low-dimensionality of the i-vector. Studies have used i-vector for various speech tasks, including speaker recognition, speaker adaption for ASR, and stress recognition etc. [18, 20–22]. The i-vector modeling is given as:

$$M = m + Tx, \tag{1}$$

where M is the GMM supervector obtained from MAP adaptation, m is the speaker, channel, emotion attribute-independent mean-vector constructed from the *universal background model* (UBM). The total variability matrix T is a low-rank projection matrix obtained from all training data by factor analysis training [23]. The i-vector is x which is a low-dimensional vector.

Table 1. Criteria to form the enrollment set for the speaker verification task, using the MSP-PODCAST database.

Criteria
CRITERION 1: Add utterances at random where the categorical emotion is "neutral" and arousal, valence, and dominance values are inside the range [3,5].
CRITERION 2: Add utterances at random where the arousal, valence, and dominance values are inside the range [3,5], regardless of the categorical emotion.
CRITERION 3: Add utterances at random where the categorical emotion is "neutral" and arousal, valence, and dominance values are inside the range [2,6].
CRITERION 4: Add utterances at random where the arousal, valence, and dominance values are inside the range [2,6], regardless of the categorical emotion.
[The range for attributes is [1-7], where 4 is neutral value.]

4.2. Mean normalized PLDA

After we extract the i-vector for each utterance, we use the mean normalized PLDA back-end to train speaker models and compute the scores. In the MSP-PODCAST corpus, the j^{th} speaker has D speech segments such as $X_j = \{x^1, ..., x^d, ..., x^D\}$. In the enrollment phase, we compute the mean of these enrollment i-vectors as the final speaker representation:

$$\bar{x}_j = \sum_{d=1}^{D} x^d, \tag{2}$$

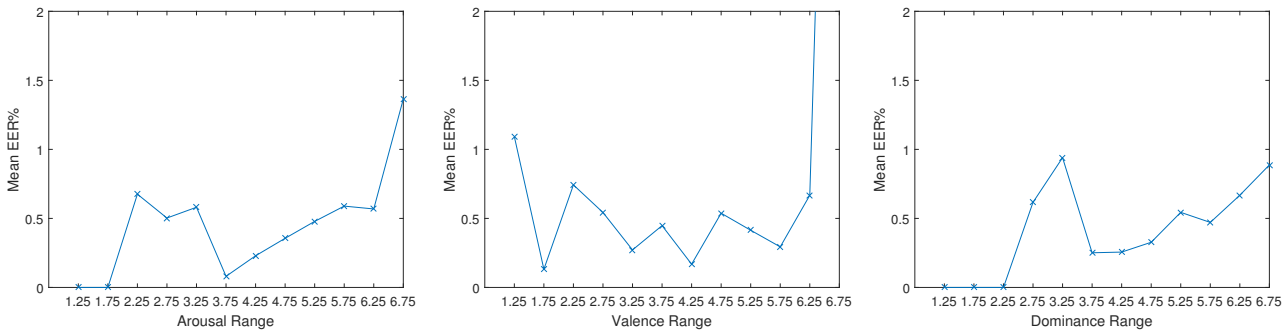
where $\bar{x}_j$ is the mean i-vector of the jth speaker for PLDA training. With the mean normalization, we believe the benefits are two-fold: a) we average out the variabilities introduced by channel, duration, context and emotional content, which helps to build a robust speaker model in the training stage [24]; b) we reduce the computational complexity when the mean normalization is applied to train PLDA. We have used this framework in NIST competitions, achieving competitive performance [24].

After detecting speech with an energy based speech activity detector, we extract a 39-dimensional MFCCs+Δ+$\Delta\Delta$ feature vector. We train a 256-mixture UBM using the training data, empirically setting the dimension of the i-vector to 200. We select *equal error rate* (EER) as the metric to evaluate the speaker verification performance.

5. EXPERIMENTAL EVALUATION

This study analyzes the performance of the speaker verification system with expressive speech. Previous studies have analyzed the performance of speaker identification systems in terms of categorical emotions (e.g., anger, happiness). We argue that the intra-class variability for a given emotion introduces artifacts in the analysis that will prevent us to identify the range of emotional content for which the performance of speaker verification system drops. For example, the acoustic properties for hot anger and cold anger are significantly different, so we expect they will affect the speaker verification performance differently. Instead, we propose to rely on the continuous attributes arousal, valence and dominance. We expect that sentences that are perceived with similar emotional attributes will have similar deviations from neutral speech, having a similar effect on the speaker verification performance.

We aim to analyze the speaker verification performance with mismatches in the training (neutral speech) and testing (emotional speech) conditions. Therefore, we only use neutral speech for enrollment. For each speaker in the MSP-PODCAST corpus, we have



(a) Arousal (b) Valence (c) Dominance

Fig. 2. Speaker verification performance in terms of arousal, valence and dominance. The figures separately analyze each emotional attribute.

over 300s of speech, including neutral and emotional data. We use 150s for enrollment, and the remaining utterances are used in the verification set. To ensure that the enrollment set consists of utterances that are mainly neutral, we consider the categorical emotion, as well as the arousal, valence, dominance scores. We follow, in order, the criteria described in Table 1 until reaching 150s per speaker. Criterion 1 selects sentences that are clearly neutral, as they are labeled with the class "neutral" and their arousal, valence and dominance scores lie inside the range [3-5] (the range for the attributes is [1-7], where 4 is the neutral value). If we still need more data to reach 150s of speech, we use criterion 2, where we select all samples with attribute values inside the range [3-5], regardless of the emotional class assigned to the sentences. The last two criteria are similar to the first two criteria, where we extend the range for the emotional attributes. With this approach, the training set has 1061 sentences (636: criterion 1; 188: criterion 2; 194: criterion 3; 43: criterion 4).

The speaker verification task consists of pairwise assessments where the problem is to decide if two speech signals belong to the same person. In a typical speaker verification system, a decision whether to reject or accept a claimed identity is made based on the enrolled speaker models. We use EER as our primary metric to evaluate how different emotional dimensional attributes influences speaker verification performance. We compute EER value for each testing utterance.

6. RESULTS

6.1. Effect of Emotional Dimensions

After conducting the speaker verification task, we analyze the EER in terms of arousal, valence and dominance scores assigned to the sentences in the test set. Figure 2 provides the results, where we group sentences with similar scores, averaging their EER. As expected, our speaker verification system drops its performance for extreme values arousal and valence and dominance. This result is particularly noticeable for high values of these attributes. Note that there are fewer samples with arousal and dominance in the range [1-2], so the results in Figures 2(a) and 2(c) for this range are less conclusive.

A limitation of results presented in Figure 2 is that the dependencies between attributes are not considered. For example, consider a sentence with an extreme score for arousal (e.g., 7) an a normal score for valence (e.g., 4). If the speaker verification performance is low, we may incorrectly conclude that neutral values of valence are detrimental for the system. We address this limitation by jointly analyzing the speaker verification performance over the arousal-valence space. We do not consider dominance for this analysis since (1) arousal and valence are the most used emotional dimensions [25] (2) dominance is usually highly correlated with arousal. Figure 3 displays the results. We created this figure by using a 0.5×0.5 window, shifting it by steps of 0.1 along each axis (e.g., valence, arousal). We average the EER of all the sentences with arousal and valence scores inside the window, assigning this value to the block. The performance is only reported at points with at least 10 or more utterances. Otherwise, the cell is left empty. The Figure 3 shows that the EER are very low for sentences with neutral values for arousal and valence (i.e., close to [4,4]). As the values of these attributes deviate from the neutral area, the average EER increases.

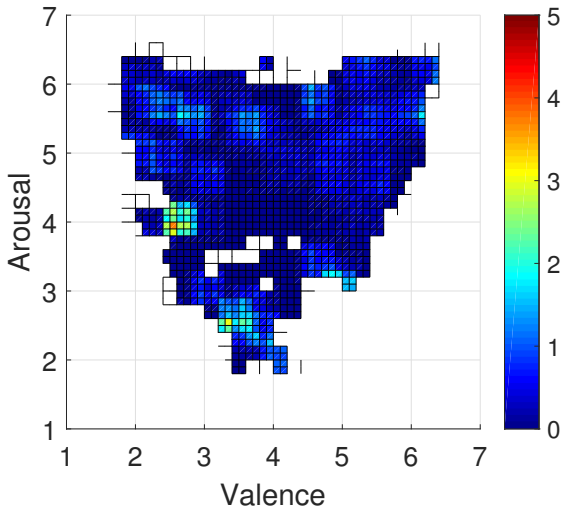


Fig. 3. Speaker verification performance in terms of arousal and valence, where each block represents average EER for sentences inside a 0.5×0.5 window centered at the block.

6.2. Effect of the Duration of the Sentence

We also evaluate the speaker verification performance in terms of the duration of the sentences. The MSP-PODCAST dataset has utterances varying between 2.75s and 11s. Since the enrollment and verification sets have utterances of varying durations, we have mismatched conditions for training the speaker verification system. First, we discretize the duration of the sentences into 10 bins. Then, we estimate the average EER for all sentences in each bin. Figure 4 gives results which show that shorter utterances decrease the speaker verification performance. Phonetic mismatch caused by short du-

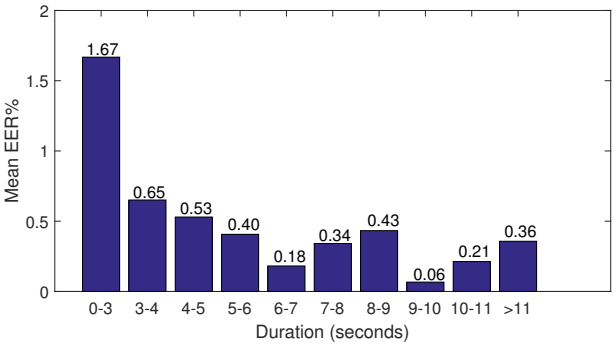


Fig. 4. Speaker verification performance as function of the length of the utterances.

ration is still very challenging in text independent speaker verification [26,27]. The error is maximum for short sentences with duration less than 3s. The mean EER drops below 0.5% for sentences longer than 5s.

6.3. Effect of Emotion for Sentences with Same Duration

To distinguish the effect of duration and emotion on speaker verification performance, we repeat the evaluation with sentences with same duration for both train and test sets. For each speaker, we accomplish this goal by creating 5s sentences as follows. First, we split longer sentences into 5s segments, assigning the emotional attribute scores of the original sentences to each of their segments. Second, segments shorter than 5s are merged according to their arousal, valence and dominance scores. For each segment, we form a 3D vector with the attribute scores, estimating the Euclidean distance between all possible pairs of segments. Then, we sort all the pairs according to the Euclidian distance, merging the segments with closer distance when their duration exceeds 5s. We keep the longest segment, and the shorter segment is cropped such that the new segment is exactly 5s. The emotional attributes for the new segment is the average of the two merged segments. We remove all the pairs where the original segments appears. We repeat this process for all the pairs with Euclidean distance below 0.5. We re-train the speaker verification system, with the 5s sentences. Given the process to merge the sentences, we do not have the categorical emotions to create the training set. Therefore, we select the neutral sentences using only criteria 2 and 4 in Table 1. Figure 5 shows the EER results for the 5s segments.

The general trends are similar to the ones with the original duration (Fig. 3). We observe darker blue regions near the neutral area, and lighter blue regions towards the upper corners. The main difference between the figures is the performance for sentences with low arousal, which now have lower EER. Areas of poor EER can be further attributed to certain specific speakers or regions without many sentences. We expect a smoother transition as we include more speakers and more testing utterances per speaker in the dataset.

6.4. Effect of Individual Speaker

We also analyze the performance of our speaker verification system per speaker. Figure 6 shows the average EER for each speaker in the database. Only four speakers out of 40 have average EER above 1%. The error tends to be distributed across most of the speakers, where 25 speakers have EER below 0.5%.

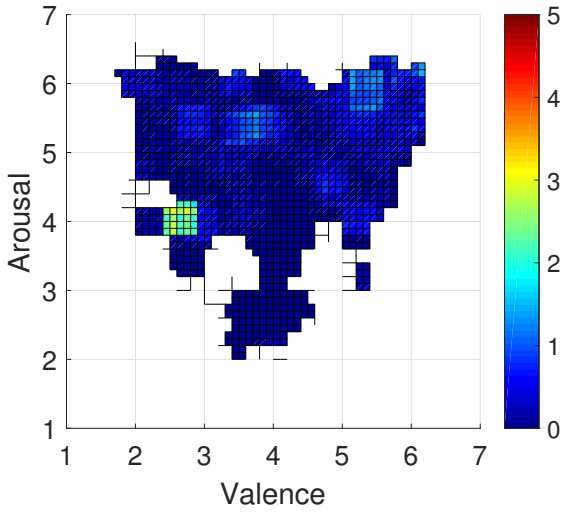


Fig. 5. The figure illustrates Speaker ID performance with respect to Arousal, Valence value of the utterance. All utterances are of equal duration of 5 sec.

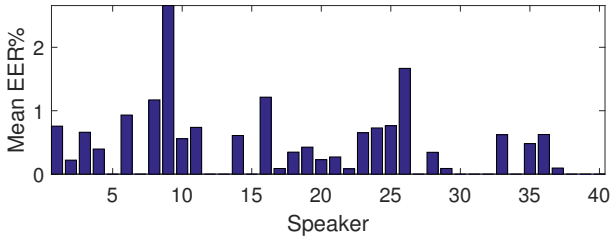


Fig. 6. The figure illustrates Speaker ID performance for different speakers in the corpus.

7. CONCLUSIONS

This paper provided a systematic analysis on the effect of emotions, described with emotional attributes (arousal, valence, dominance), on speaker verification performance. The study relies on a portion of the MSP-PODCAST corpus consisting of 40 speakers with at least 5 minutes of naturalistic speech for every speaker. The naturalistic recordings and the number of speakers considered in this analysis overcome limitations of previous studies. By training the models with mainly neutral speech, we evaluate the speaker verification performance in mismatched conditions as function of the perceived arousal, valence and dominance scores of the sentences. The results demonstrated that the performance drops for sentences with extreme values for these attributes. We also analyzed the performance in terms of the durations of the sentences, showing that the length of the utterances is an important factor for verification systems. By artificially merging and splitting sentences with similar emotions, we demonstrated that the reported trends are also observed for sentences with same duration (5s).

The analysis in this study opens interesting research questions on affective computing in the area of speaker verification. We are planning to estimate emotional attributes using machine learning frameworks. The estimated values for arousal, valence and dominance can serve as a tool to predict the reliability of speaker verification system in the presence of expressive speech.

8. REFERENCES

[1] C. Busso, S. Lee, and S.S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[2] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[3] M. Abdelwahab and C. Busso, "Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, South Lake Tahoe, CA, USA, December 2014, pp. 472–477.

[4] T. Athanaselis, S. Bakamidis, I. Dologlu, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: Clarifying the numbers and enhancing performance," *Neural Networks*, vol. 18, no. 4, pp. 437–444, May 2005.

[5] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *ISCA Speech Prosody 2006*, Dresden, Germany, May 2006, ISCA.

[6] I. Shahin, "Speaker identification in emotional talking environments based on CSPHMM2s," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1652–1659, August 2013.

[7] H. Bao, M.X. Xu, and T.F. Zheng, "Emotion attribute projection for speaker recognition on emotional speech," in *Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 758–761.

[8] D. Li, Y. Yang, Z. Wu, and T. Wu, "Emotion-state conversion for speaker recognition," in *Affective Computing and Intelligent Interaction (ACII 2005)*, J. Tao, T. Tan, and R.W. Picard, Eds., vol. 3784 of *Lecture Notes in Computer Science*, pp. 403–410. Springer Berlin Heidelberg, Beijing, China, October 2005.

[9] M. V. Ghiurcau, 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 2011)*, Prague, Czech Republic, May 2011, pp. 4944–4947.

[10] W. Wu, T.F. Zheng, M.X. Xu, and H. Bao, "Study on speaker verification on emotional speech," in *International Conference on Spoken Language (ICSLP 2006)*, Pittsburgh, PA, USA, September 2006, pp. 2102–2105.

[11] Z. Wu, D. Li, and Y. Yang, "Rules based feature modification for affective speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, Toulouse, France, May 2006, vol. 1, pp. 661–664.

[12] S.R. Krothapalli, 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*, vol. 15, no. 3, pp. 335–349, September 2012.

[13] I. Shahin, "Speaker identification in emotional talking environments using both gender and emotion cues," in *International Conference on Communications, Signal Processing, and their Applications (ICCSPA 2013)*, Sharjah, United Arab Emirates, February 2013, pp. 1–6.

[14] D. Li and Y. Yang, "Emotional speech clustering based robust speaker recognition system," in *International Congress on Image and Signal Processing (CISP 2009)*, Tianjin, China, October 2009, pp. 1–5.

[15] S.G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, June 2012.

[16] M.V. Ghiurcau, C. Rusu, and J. Astola, "Speaker recognition in an emotional environment," in *Signal Processing and Applied Mathematics for Electronics and Communications (SPAMEC 2011)*, Cluj-Napoca, Romania, August 2011, pp. 81–84.

[17] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[18] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[19] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[20] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J.H.L. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2854–2857.

[21] C. Zhang, G. Liu, C. Yu, and J. H.L. Hansen, "I-vector based physical task stress detection with different fusion strategies," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2689–2693.

[22] C. Zhang, S. Ranjan, M.K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H.L Hansen, "Joint information from nonlinear and linear features for spoofing detection: an i-vector/dnn based approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5035–5039.

[23] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, May 2005.

[24] G. Liu, T. Hasan, H. Boril, and J.H.L Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 7755–7759.

[25] J.A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.

[26] T. Hasan, R. Saeidi, J. H.L. Hansen, and D. A van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 7663–7667.

[27] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 7649–7653.