DETECTING THE INSTANT OF EMOTION CHANGE FROM SPEECH USING A MARTINGALE FRAMEWORK

Zhaocheng Huang[1,2] and Julien Epps[1,2]

[1]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia
zhaocheng.huang@student.unsw.edu.au, j.epps@unsw.edu.au

ABSTRACT

Towards a better understanding of emotion in speech, it is important to understand how emotion changes and when it changes. Recognizing emotions using pre-segmented speech utterances results in a loss in continuity of emotions and does not provide insights into emotion changes. In this paper, we propose an investigation into emotion change detection from the perspective of exchangeability of data points observed sequentially using a martingale framework. Within the framework, a per-frame GMM likelihood based approach is proposed as a measure of strangeness from a particular emotion class. Experimental results on the IEMOCAP database demonstrate that the proposed martingale framework offers significant improvements over the baseline GLR method for detecting emotion changes not only between neutral and emotional speech, but also between positive and negative classes along the arousal and valence emotion dimensions.

Index Terms **—** Emotion change detection, Gaussian mixture model, Martingale, Exchangeability

1. INTRODUCTION

When speakers engage in human-computer interaction during which their emotions are recognized from behavioral signals, it is desirable that the system can detect changes in emotions as they occur, so that it can react correspondingly. Emotion recognition research to date has mainly focused on classifying or predicting from pre-segmented speech signals (e.g. on a file-by-file basis) [1], [2], [3], which lacks realism and does not provide insight into emotion changes. As emotions are essentially responses or reactions to external stimuli, understanding emotion changes might help understanding the external environment, such as what is happening to the speaker, or what triggers these emotions. These demands motivate research aiming to detect emotion changes in time regarding emotion categories [4] and emotion dimensions [5]. However, emotion change detection remains challenging and far from being used in applications, due to variability (e.g. phonetic and speaker variability) in speech, and the subjective nature of emotions (e.g. neutral and sadness are confusable in speech).

Although emotion change detection is an understudied research area, change detection has been a long-standing problem for example in speaker change detection [6], concept drift detection [7] and video shot change detection [8]. Speaker change detection methods have recently been investigated for emotion change [4]. However many of these methods seem to either be easily affected by variability or depend on the availability of large databases [9], [10], [11]. Because of this, methods from the more generic problem of statistical change detection are considered, which motivates investigation of martingales. Unlike most change detection methods using large sliding windows, martingales have been proposed for detecting changes in streaming data and make decisions on-the-fly by testing exchangeability (refer to section 3.1) of sequentially observed data points [8], [12]. This opens a possibility for an alternative framework for emotion change detection with an improvement in temporal resolution [7].

In this paper, the problem of localizing emotion change points in time was investigated from the perspective of testing exchangeability using a martingale framework, where data points (frame-based features) from speech are observed one by one. This method potentially offers higher temporal resolution than using large sliding windows. Moreover, emotional models may be helpful to reduce effects of phonetic variability compared with methods that require no prior knowledge of emotions.

2. RELATED WORK

Although emotion change detection might be important for a better understanding of emotions, there have been only a few studies that aim to detect the instant of emotion changes [4], [5]. Lade proposed an adaptive temporal topic model that captures the temporal information for localizing the time when huge changes in emotion dimensions occur [5]. In our previous work [4], GMM based methods with and without prior knowledge of emotion were proposed to detect any emotion changes among four emotions. However, an important emotion change of interest is between neutral and emotional speech. Also, a focus on arousal and valence would allow a move away from pre-defined categories to more generic descriptors of the emotion space. There are currently many papers focusing only on +/- arousal and +/- valence, e.g. for cross-corpus comparisons [13].

An intuitive question with emotion change detection is how it compares with performing emotion recognition in real time and finding resultant changes. However, the accuracy of speech based emotion recognition [1], [2], [3], [14], [15], [16], [17], especially for naturalistic and semi-naturalistic databases, remains unsatisfactory for being used for the purpose for change detection.

Detecting emotion changes is somewhat analogous to speaker change detection, e.g. [6], [18], [19], [20]. However, applying these methods into an emotion change detection task remains problematic because of the phonetic and speaker variability embedded in emotional speech and the complex nature of emotion (e.g. a person might experience more than one emotion at a time). Therefore, our attentions have been focused on a more generic change-detection problem, where the martingale-based methods based on testing exchangeability (discussed further in section 3.1) are applied [7], [8]. The idea of exchangeability was

introduced by [12], and applied for change detection by [8]. It has been widely used in image processing [21]. However, there has very little work done in speech apart from [22], in which the martingale method was used for detecting speech rate changes. Despite the difference in context, their work showed that testing exchangeability could be effectively used in speech processing. After introducing the original martingale framework as well as analyzing its potential drawbacks for detecting emotion changes (section 3.2), a modified framework is proposed (section 3.3).

3. A MARTINGALE FRAMEWORK

3.1 Exchangeability and Martingale

By definition [12], a sequence of random variables $\{x_1, x_2, \dots x_n\}$ is exchangeable if their joint distribution remains unchanged regardless of any permutation π of $\{1, \dots, n\}$, namely:

$$p(x_1, x_2, \dots x_n) = p(x_{\pi(1)}, x_{\pi(2)}, \dots x_{\pi(n)}) \quad (1)$$

An example of exchangeability is the selection of balls in sequence without replacement from an urn where there are only uniquely numbered red balls. In this case, the joint probability of choosing red balls remains invariant. Consider the case that a number of red balls are selected up to time *t*, after which one starts to select balls from another urn for which the probability of a red ball is $p(red) < 1$. Then the selected ball sequence becomes less exchangeable, as the joint distribution of selecting a red ball is no longer 1. The changes in model or distribution undermine exchangeability.

Given a sequence of random variables $\boldsymbol{X}_i: \{x_1, x_2, \dots x_i\}$, where $\boldsymbol{X}_i$ denotes all random variables from 1 to *i*, if M_i is a measurable function of $\boldsymbol{X}_i$ and $E(|M_i|) < \infty$, then $\{M_i : 0 \leq i \leq \infty\}$ is a *Martingale* process once it satisfies [8], [23]:

$$E(M_{n+1}|\boldsymbol{X}_n) = M_n \quad (2)$$

Further, the terms *Submartingale* and *Supermartingale* can be defined respectively as:

$$E(M_{n+1}|\boldsymbol{X}_n) \geq M_n \quad (3)$$

$$E(M_{n+1}|\boldsymbol{X}_n) \leq M_n \quad (4)$$

For change detection, the martingale value M_i measures the confidence of rejecting the null hypothesis of exchangeability. Combining exchangeability and martingales, a family of *Randomized Power Martingales* with initial value $M_0 = 1$, was proposed by [12]:

$$M_n^{(\varepsilon)} = \prod_{i=1}^{n} \varepsilon p_i^{\varepsilon - 1} \quad (5)$$

where p_i can be seen as a measure of the exchangeability and will be discussed in detail in the following section. $\varepsilon \in [0,1]$ controls the threshold for transitions between the supermartingale and submartingale. From equations (2) and (5), $M_n^{(\varepsilon)}$ is a martingale process once

$$\varepsilon p_i^{\varepsilon - 1} = 1 \quad (6)$$

$$p_i = e^{\frac{ln(\varepsilon)}{1-\varepsilon}} \quad (7)$$

Also according to equation (3), (4) and (7), $M_n^{(\varepsilon)}$ becomes a supermartingale when $p_i > e^{\frac{ln(\varepsilon)}{1-\varepsilon}}$, whereas $M_n^{(\varepsilon)}$ becomes a submartingale when $p_i < e^{\frac{ln(\varepsilon)}{1-\varepsilon}}$. A submartingale occurs when the data points observed are no longer exchangeable and $M_n^{(\varepsilon)}$ starts increasing. Once $M_n^{(\varepsilon)}$ is larger than a defined threshold, the null hypothesis of no change is rejected, as seen in Figure 1(a).

3.2 Martingale Framework for Change Detection

Based on exchangeability testing, it is crucial to calculate *p* values that are representative of exchangeability. There are two main steps: strangeness and *p* value calculation. Strangeness measures how different a data point is from others with respect to a model λ, which can be expressed for a data point $\boldsymbol{x}_n$ as

$$s_n = f(\boldsymbol{x}_n, \lambda) \quad (8)$$

The larger s_n is, the less likely the data point x_n comes from the model λ. Then the corresponding *p* value of s_n can be calculated as follow [8]:

$$p_n(X_n, \theta_n) = \frac{\#\{i: s_i > s_n\} + \theta_n \#\{i: s_i = s_n\}}{n} \quad (9)$$

Where #{} is the number of elements satisfying the bracketed condition and $\theta_n \in [0,1]$ is a random number.

If there is no change, the observed data points are all from the same model λ, which implies similarity and exchangeability of $s_1, \dots, s_n$. Accordingly, the *p* values are uniformly distributed on [0, 1], and $E(p_n) = 0.5 > e^{\frac{\ln(\varepsilon)}{1-\varepsilon}} \in [0, 0.3679]$, depending on ε. According to (5), $M_n^{(\varepsilon)}$ is preferably a supermartingale and decreases with some fluctuations due to the fact that p_i is random. Once observed data points are not exchangeable, strangeness s_n becomes larger and in turn p_n in (9) has a smaller value. Accordingly, $M_n^{(\varepsilon)}$ increases until reaching a predefined threshold over which a change is detected. Once a change is detected, $M_{n+1}^{(\varepsilon)}$ is reset to 1 and the detection restarts based on a new model trained using recent samples (50 samples in Figure 1 (a)).

This martingale framework has three drawbacks. The first is that enough data points are needed for increasing $M_n^{(\varepsilon)}$ until over the threshold, which leads to a delay, ranging from 100 to 200 data points [8]. The second one is that this method cannot handle change detection within a large time period because $M_n^{(\varepsilon)}$ will be fairly small if the exchangeability remains for a long time. Preliminary experiments also showed that when this martingale method was directly applied in the emotion change detection context, its performance degraded possibly because a martingale using frame-based acoustic features may detect changes in phonemes [22]. Thus, a modified martingale-based method using emotion model is proposed in the following section.

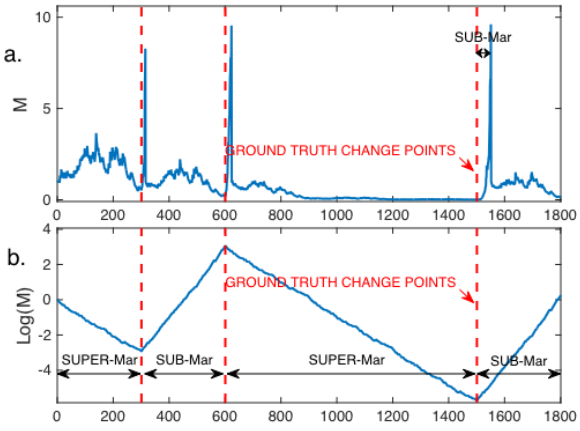

Figure 1: *Comparison of original martingale (a) and proposed martingale (b). In this simplified example, samples are randomly generated from Gaussian distributions of two classes with mean shift of 1. The strangeness measure is the negative log likelihood of*

a single Gaussian distribution trained using most recent 50 samples after a change is detected for (a) and only the first 50 samples for (b). The delay in (a) is the distance between the ground truth change points and the peaks of martingale values.

3.3 Proposed Martingale Framework for Emotion Change Detection

To resolve the aforementioned problems, a p value thresholding method is proposed which enforces floor and ceiling p values, and poses emotion change detection as a martingale turning point detection problem, in which peaks and troughs are the indicators of emotion changes, as seen in Figure 1(b):

1) Extract frame-level d-dimensional acoustic features $\boldsymbol{x}_n$ from speech, with frame shifts of 10 ms.
2) Calculate a strangeness value s_n, which is the negative of the log likelihood given the feature vector $\boldsymbol{x}_n$ and a Gaussian mixture model $\lambda(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{C})$.

$$s_n = f\left(\boldsymbol{x}_n, \lambda(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{C})\right) = -log\left(\sum_{i=1}^{m} \omega_i \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{C}_i|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x_n-\mu_i)^T \boldsymbol{C}_i^{-1}(x_n-\mu_i)\right)}\right) \quad (10)$$

3) Calculate p values based on s_n,

$$p_n = \begin{cases} p^{sub}, & s_n \geq S \\ p^{super}, & s_n < S \end{cases} \quad (11)$$

where $p^{sub} \in [0, e^{\frac{ln(\varepsilon)}{1-\varepsilon}})$ and $p^{super} \in (e^{\frac{ln(\varepsilon)}{1-\varepsilon}}, 1]$ are the parameters that activate the *Randomized Power Martingales* into submartingale and supermartingale respectively. S, a threshold for exchangeability, is an important parameter in the sense that a high S tolerates some data points that are less likely from model $\lambda(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{C})$, whereas a small S rejects some data points that are likely from model $\lambda(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{C})$. Both cases lead to unreliable transition between submartingale and supermartingale using (11), which in turn practically give rise to $log(M)$ characteristics with incorrect turning points. To address this problem, S is calculated as a trade-off between distributions of two classes:

$$S = \frac{(\boldsymbol{S}_Q^1 + \boldsymbol{S}_{100-Q}^2)}{2} \quad (12)$$

Where $\boldsymbol{S}_Q^1$ denotes the Q % percentile of all the strangeness values of class 1, estimated using ground truth from other speakers. One advantage of estimating S using (12) is that this can offer good separation between two classes, avoiding relatively large and small S to make sure that transition between submartingale and supermartingale occurs when there is a change.

4) Calculate randomized power martingale $\log(M_n)$ using (5).
5) Detect turning points using two-pass linear regression.

$$k_{past}^{N1} * k_{future}^{N1} < 0 \quad (13)$$

$$k_{past}^{N2} * k_{future}^{N2} < 0 \quad (14)$$

Where $N1$, $N2$ are the number of samples used to fit linear regressors to $\log(M_n)$ for calculating the slope k. The approach rejects the null hypothesis of no change once (13) and (14) hold, and detects a change. The reason behind the two-pass linear regression is because only using a small N can detect turning points more precisely in time but is vulnerable to noise, whereas only using a large N is robust to noise but leads to large delay.

Compared with the original martingale framework, the proposed approach formulates a turning point detection problem, which requires no threshold and reduces the delay (as seen in Figure 1). Moreover, the proposed method can handle non-change for a long time period, during which case $log(M)$ simply continues to increase or decrease until changing the direction when there is a change. Finally yet importantly, emotion model λ seems to improve the handling of phonetic variability.

4. EVALUATION

4.1. Database

In this paper, we aim to detect changes between emotion categories (neutral vs emotional) as well as in dimensions (positive vs negative in arousal and valence). The IEMOCAP database, which comprises 12 hours of emotional speech from 10 speakers, was used. As ground truth for change points were not provided in the IEMOCAP database, a new database was constructed using the following scheme, as per [4]:

- Concatenate same-speaker emotional utterances to generate ground truth change points
- Further modify the database by removing small utterances (3s for emotion categories and 7s for emotion dimensions) and repeat the above step

To investigate emotion categories, utterances of neutral, anger, sadness, happiness and excitation with majority consensus were selected. The latter four emotions were then merged into an "emotional" class (EMO). To investigate emotion dimensions, initially all utterances were selected (10039 utterances). Then numerical ratings were z normalized and thresholded into positive and negative respectively using thresholds of ± 0.7, similar to [16], resulting in different datasets for arousal and valence (Table 1).

Table 1. *Partitions from IEMOCAP used for experiment. The voicing probability thresholds were the default 0.55 for MFCCs and 0.7 for eGeMAPS features [24], leading to different partitions for the two sets of features.*

	MFCCs		#changes	eGeMAPS		#changes
	+/emo	-/neu		+/emo	-/neu	
EMO	3785	1698	186	3789	1697	224
Arousal	1847	3063	123	1844	3073	207
Valence	2808	3195	169	2807	3196	196

4.2. Experimental Settings

Two sets of frame-level acoustic features were extracted using the *openSMILE* toolkit [25]. The first set was 13 MFCCs and their first derivatives. The second set is the 28-dimension extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [10], a knowledge-based set of features, which are effective in emotion recognition tasks. Leave-one-speaker-out 16 mixture GMMs models (a trade-off between computational complexity and detailed description of emotions) for neutral or negative dimension (arousal/valence) classes were trained for the strangeness calculation (10). The threshold S was estimated from 9 speakers using the GMM model. The baseline is the Generalized Likelihood Ratio (GLR) method [4] using a sliding dual windowing framework with one Gaussian (diagonal covariance) and window sizes of 1 second (for emotion categories) and 3.5 seconds (for emotion dimensions), based on the best results obtained using different window sizes. Within the proposed framework, there are a number of parameters such as ε, p^{sub}, p^{super}, Q, $N1$ and $N2$. Among these, ε, p^{sub} and p^{super} control how fast the $log(M)$ increases and decreases, which does not affect the turning points and therefore performance is less sensitive to their choice. ε was set to 0.92 in (5) according to [8]. p^{sub} and p^{super}, the parameters ensuring the randomized power martingale becomes a submartingale and a supermartingale, were set to 0.25 and 0.5 in

(11) respectively. The turning points are sensitive to choices of S. However, the method (12) proposed for calculating S offers a relatively good separation of two classes. Q for calculating threshold for strangeness values was set to 70% and 50% in (12) for comparison. This is because the proposed martingale framework requires no threshold, and changing Q leads to a different S and in turn different $log(M)$ characteristics. In the two-pass linear regression, choices of $N1$ and $N2$ are related to accuracies of detecting turning points. $N1$ and $N2$ were set to 10 and 60 empirically. The tolerance region for change detection was set to 1 second.

4.3. Results

4.3.1 Emotion Change Detection for Emotion categories and dimensions

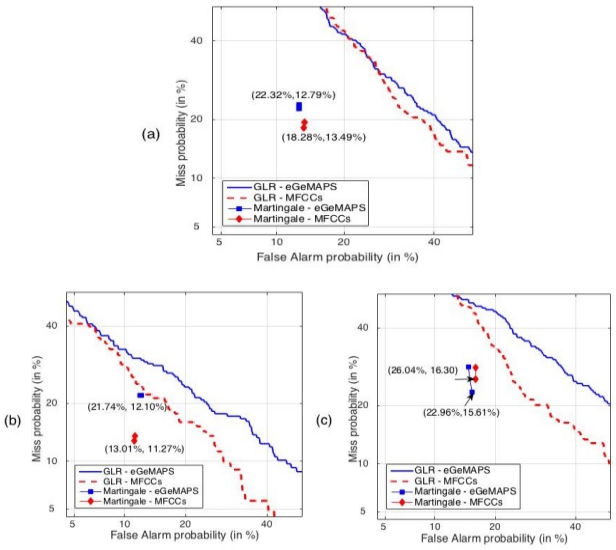

Figure 1: *DET curves for the proposed martingale method and the baseline GLR method [4] using MFCCs and eGeMAPS feature set for detecting changes (a) between neutral and emotional speech; (b) between positive and negative arousal; and (c) between positive and negative valence. Q was set to 50 and 70, which leads to two operating points for the martingale-based method.*

Firstly, experiment was conducted comparing the proposed martingale and the baseline GLR method [4] for *three* different tasks, namely emotion change detection for neutral and emotional speech; positive and negative in arousal and valence. Regression based methods proposed in [5] are considered unsuitable for direct comparison. Consistent significant improvements over the baseline can be seen using the proposed martingale method. Large differences in performance were seen when detecting positive and negative valence for different Qs. The GLR method, which requires no prior knowledge of emotion, is not very effective for the neutral vs emotional change detection, possibly due to the fact that there are more salient emotion changes (e.g. change between anger and sadness within emotional speech), as well as phonetic variability. The martingale method, which tests exchangeability with respect to emotion model, showed that inclusion of prior knowledge of emotion is helpful. MFCCs were more effective in the GLR method, and it was found that within the martingale framework, MFCCs features are advantagous for arousal change detection, whereas interestingly the eGeMAPS feature set have a better performance in valence change detection.

4.3.2 Tolerance Region Duration: a Trade-off between Temporal Resolution and Detection Accuracy

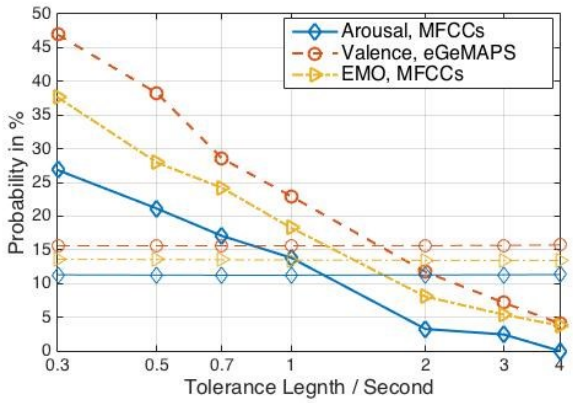

Figure 2: *Miss Detection Probability (thicker lines) and False Alarm Probability (thinner lines) vs tolerance region lengths for the three tasks using the feature set that provided the best performance within the proposed martingale framework.*

A tolerance region is essentially a temporal window around which a ground truth change point occurs [4]. Within this window, any detected changes are regarded as a correct detection. It allows a trade-off between temporal resolution and detection accuracy in change detection tasks, as seen in Figure 3. With a tolerance of 3s, all three tasks have MD probability lower than 10%, whereas FA is constant for all tolerance lengths, because the proposed method is based on turning point detection.

5. CONCLUSIONS

This paper has presented a martingale-based framework for emotion change detection by testing exchangeability. This framework poses change detection as a turning point detection problem, and a two-pass linear regression method was used to detect peaks and troughs. This is advantageous in terms of a lower delay, capacity to handle non-change over a relatively long time, and robustness to potential variability. Exchangeability of each frame was measured by using the negation of log likelihood of a GMM model. Experimental results demonstrate that performances were boosted over the baseline for both emotion categories and dimensions by using the proposed framework.

Two limitations of this work are the limited possible operating points of the proposed approach and the database, in which all emotional utterances were concatenated and further modified for each speaker to create changes. Future work includes testing the proposed framework in more realistic databases. Moreover, novelty detection methods might be a good fit for this framework, because they require only one model to be effective. Rather than observing data points only at frame-level, turn-level functionals, successfully applied in emotion recognition previously to address phonetic variability, can potentially improve system performance.

6. ACKNOWLEDGEMENTS

This work was partly funded by the US Army International Technology Center (Pacific). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

7. REFERENCES

[1] Lee C.-C., Mower E., Busso C., Lee S., and Narayanan S., "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication,* vol. 53, pp. 1162-1171, 2011.

[2] Schuller B., Batliner A., Steidl S., and Seppi D., "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication,* vol. 53, pp. 1062-1087, 2011.

[3] Dumouchel P., Dehak N., Attabi Y., Dehak R., and Boufaden N., "Cepstral and long-term features for emotion recognition," in *INTERSPEECH*, 2009, pp. 344-347.

[4] Huang Z., Epps J., and Ambikairajah E., "An Investigation of Emotion Change Detection from Speech," in *INTERSPEECH*, 2015.

[5] Lade P., Balasubramanian V. N., Venkateswara H., and Panchanathan S., "Detection of changes in human affect dimensions using an Adaptive Temporal Topic model," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, 2013, pp. 1-6.

[6] Miro X. A., Bozonnet S., Evans N., Fredouille C., Friedland G., and Vinyals O., "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 20, pp. 356-370, 2012.

[7] Pears R. S., Sripirakas and Koh Y.-S., "Detecting concept change in dynamic data streams," *Machine Learning,* vol. 97, pp. 259-293, 2014.

[8] Ho S.-S. and Wechsler H., "A martingale framework for detecting changes in data streams by testing exchangeability," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 32, pp. 2113-2127, 2010.

[9] Siegler M. A., Jain U., Raj B., and Stern R. M., "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Broadcast News Workshop*, 1997, p. 11.

[10] Kenny P., Reynolds D., and Castaldo F., "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of,* vol. 4, pp. 1059-1070, 2010.

[11] Silovsky Jan P. J., "Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4193-4196.

[12] Vovk V., Nouretdinov I., and Gammerman A., "Testing Exchangeability On-Line," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, p. pages.

[13] Schuller B., Vlasenko B., Eyben F., and Martin W., "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 470-476.

[14] Jin Q., Li C., Chen S., and Wu H., "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, Brisbane, Australia, 2015, pp. 4749-4753.

[15] Kockmann M. and Burget L., "Application of speaker-and language identification state-of-the-art techniques for emotion recognition," *Speech Communication,* vol. 53, pp. 1172-1185, 2011.

[16] Abdelwahab M. and Busso C., "Supervised Domain Adaptation for emotion recognition from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, Brisbane, Australia, 2015.

[17] Chen L., Mao X., Xue Y., and Cheng L. L., "Speech emotion recognition: Features and classification models," *Digital Signal Processing,* vol. 22, pp. 1154-1160, 2012.

[18] Chen S. and Gopalakrishnan P., "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, p. 8.

[19] Gish H., Siu M.-H., and Rohlicek R., "Segregation of speakers for speech recognition and speaker identification," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1991, pp. 873-876.

[20] Fergani B., Davy M., and Houacine A., "Speaker diarization using one-class support vector machines," *Speech Communication,* vol. 50, pp. 355-365, 2008.

[21] Mozafari N., Hashemi S., and Hamzeh A., "A precise statistical approach for concept change detection in unlabeled data streams," *Computers & Mathematics with Applications,* vol. 62, pp. 1655-1669, 2011.

[22] Yasuda H. and Kudo M., "Speech rate change detection in martingale framework," in *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*, 2012, pp. 859-864.

[23] Basseville M. and Nikiforov I. V., *Detection of abrupt changes: theory and application* vol. 104: Prentice Hall Englewood Cliffs, 1993.

[24] Eyben F., Scherer K., Schuller B., Sundberg J., e E. A., Busso C.*, et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *Affective Computing, IEEE Transactions on,* 2015.

[25] Eyben F., Weninger F., Gross F., and Schuller B., "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835-838.