Using Audio-Derived Affective Offset to Enhance TV Recommendation

Sven Ewan Shepstone, *Member, IEEE*, Zheng-Hua Tan, *Senior Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—This paper introduces the concept of affective offset, which is the difference between a user's perceived affective state and the affective annotation of the content they wish to see. We show how this affective offset can be used within a framework for providing recommendations for TV programs. First a user's mood profile is determined using 12-class audio-based emotion classifications. An initial TV content item is then displayed to the user based on the extracted mood profile. The user has the option to either accept the recommendation, or to critique the item once or several times, by navigating the emotion space to request an alternative match. The final match is then compared to the initial match, in terms of the difference in the items' affective parameterization. This offset is then utilized in future recommendation sessions. The system was evaluated by eliciting three different moods in 22 separate users and examining the influence of applying affective offset to the users' sessions. Results show that, in the case when affective offset was applied, better user satisfaction was achieved: the average ratings went from 7.80 up to 8.65, with an average decrease in the number of critiquing cycles which went from 29.53 down to 14.39.**

***Index Terms*—Affective offset, circumplex model of affect, critique-based recommenders, emotions, EPG, moods.**

I. Introduction

EVEN with the steady increase of on-demand services such as Netflix and HBO,[1] broadcast TV is still firmly entrenched in the home. It is typically the place where the local news and programming is to be found, where many consumers would be reluctant to part with. It is easy to use - turn on the TV, find a channel and watch. Since the consumer does not take part in the selection of the program lineup, recommendations can be serendipitous, something that customers value. From the provider-side there have been substantial investments in satellite, terrestrial and cable networks, and they want to see the best return on investment. Thus it is anticipated that broadcast TV will not be going away any time soon.

The Electronic Program Guide (EPG) is today an integrated part of most home television sets and set top boxes, and is still the predominant method when it comes to navigating both currently-showing and up-and-coming TV programs in the broadcast realm. It is typically presented in a grid-like fashion, with channels down and programs across the grid. However, while the EPG does provide the consumer some assistance, there can still be an overwhelming amount of content to choose from. Not only is the currently-airing program of interest, but also future programs that the consumer may wish to record or be reminded about. To illustrate: over a three-hour period with 30 available channels, with a program length of 30 minutes (typical during prime-time viewing), there are 180 programs to choose from, making an informed decision difficult.

In order to recommend something personal, a user profile is needed. The user profile data can be collected explicitly, e.g. by requesting users to supply data, or implicitly, through usage patterns. Matching of the user profile to the potentially recommendable content of interest can take place at two levels. At the cognitive level, semantic information such as content descriptors, e.g. genre or user ratings are utilized. The affective level on the other hand deals with the emotional context of the user, and how this relates to the content. The notion of cognitive and affective levels is not a new idea, and has been proposed before in the context of video content retrieval [1].

One area that has received little attention, in the context of recommending content within the EPG framework, is using the user's direct audio environment to extract profile information that can be used to make recommendations. State-of-the-art speaker recognition methods have made it substantially more feasible to extract information about the users, such as their age and gender [2] or emotions [3], using models built upon a text-independent speaker recognition framework.

In this paper we propose a novel framework that takes into account users' audio derived moods to provide the most relevant TV channel recommendation to them. A state-of-the-art audio classifier classifies users' speech into individual emotions, which contribute ultimately to their mood. Since, for a given mood, two separate users might have different ideas of what would be applicable to watch, we do not expect them to find the initially recommended item immediately appealing [4]. Users are therefore given the possibility to critique the item by navigating the emotion space of all candidate items to find a more

Manuscript received January 17, 2014; revised May 03, 2014; accepted July 04, 2014. Date of publication July 10, 2014; date of current version October 13, 2014. This work was supported by Bang and Olufsen A/S and by Denmark and the Danish Ministry of Science, Innovation, and Higher Education under Grant 12-122802. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ali C. Begen.

S. E. Shepstone is with Bang and Olufsen A/S, Struer 7600, Denmark (e-mail: ssh@bang-olufsen.dk).

Z.-H. Tan and S. H. Jensen are with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark (e-mail: zt@es.aau.dk; shj@es.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2337845

[1]Both Netflix and HBO offer broadband delivery over IP networks (HBO is traditionally a cable and satellite TV provider, but also offers broadband delivery through their On Demand service).

suitable item, should they wish to do so. To quantify the difference between the initial item and finally selected item, we model what we call the *affective offset* between the items. The novelty lies in leveraging this affective offset to provide system adjustments in such a way that future recommendations are more tailored to the individual person.

This paper is organized as follows: Section II starts with a discussion on psychological emotion theory, and how this relates to the proposed framework. Section III gives an overview of emotion detection in speech. The following section then introduces critique-based recommender systems. Section V presents the recommendation framework, discussing aspects relating to mood detection, critiquing and affective offset. Section VI presents our experimental work and the following section discusses the findings. The final section concludes the paper, and provides some recommendations for future work.

II. Moods and Emotions

Since moods cannot be measured in the same way as emotions, there is still a lot that is not yet understood about moods. We know however, that emotions, which are more transient in nature, give rise to moods and that certain emotions, such as anger, can cause one to be in a bad mood for a longer period. If certain emotions are experienced strongly and often enough over a given time period, they might eventually give rise to moods, e.g. a continuous sequence of events that cause irritation might lead one to be in a bad mood. A person with a propensity for being in bad moods, might more easily be triggered into becoming angry. While there is no agreement in the literature on how long a mood lasts, it is generally understood that moods last longer than emotions [5].

There is a difference between the mood a person is in and the *pervasive mood* of the content item they might want to see [4]. Mood management theory suggests that people will generally follow their hedonistic desires [6], meaning for example, that somebody in a bad mood might want to watch good mood content to *repair* their negative affective state.

Since we cannot measure mood directly, we concern ourselves with the actual emotions, and how they might be used to determine an entry mood for the system. A person's emotional state can be acquired either explicitly or implicitly. Due to problems seen with explicit acquisition of emotions [4], it has been suggested that they be collected implicitly. Of all the induction methodologies available for obtaining the emotional state, speech is the cheapest and most non-intrusive method.

While the modeling of emotions themselves has always been a very controversial topic [7], the most prominent model used is the dimensional approach, which is based on separate *valence*, *arousal* and *dominance* dimensions, where any emotional state can be represented as a linear combination of these three basic dimensions. Recent studies show that the valence and arousal axes account for most of the variance [1] and that these are typically the two prominent dimensions used in digital systems. We follow in the same vein. In the *VA* space, valence is more commonly referred to as the pleasantness of the emotion, whereas arousal refers to the actual intensity.

The well-known dimensional model known as the Circumplex Model of Affect [8] which is also based on valence and arousal shows how emotional states exhibit a very particular ordering around the periphery of a circle. Emotions that are close in nature are adjacent to one another, whereas bipolar emotions are situated on opposites sides of the circle. Furthermore the emotional states are not distributed evenly around the circle, and some states lie closer to each other on the circumplex than others. The location of these affective states has been determined using empirical data from psychological studies. Each location is expressed in degrees going counterclockwise around the circle, starting at 0° from the positive valence axis. While there is general agreement on the location of the emotional states, several studies have concluded different exact locations, and recent updates to these models have been made using more stringent statistical models [9].

Not only are there different interpretations of the locations of these states, but very interestingly, the very orientation of the valence-arousal axes has been debated [10]. Some studies have proposed shifting the axes, for example, by orienting them at a 45° angle, or by placing the axes where the emotions are most densely clustered.

While the valence-arousal model is well suited to Human Computer Interaction (HCI) applications, distinct emotion categories, as used in most emotion speech databases today, are not. It can therefore be difficult to relate these fixed categories to the valence arousal *VA* space. Furthermore, labeling of elicited emotions with universal labels has come under scrutiny [7], where it has been postulated that the actual felt emotions, for example, as shown in physiological readings, such as increased heart rate, might not be the same as the emotion labels themselves. This has especially been demonstrated with studies from non-western cultures. A previous work has for example looked at mapping from the *VA* space to distinct emotion categories, using clustering with density estimation [11]. However not only is this more in the context of affective video labeling, but it relies on an intuitive interpretation of what emotion each cluster is assigned to. This can be particular tricky for emotions very close to one other, and where the ordering of the clusters might change, such as in the case for the emotions fear and anger.

This study uses the Circumplex Model of Affect to model the fixed emotions and the *VA* space to model the content items. The Circumplex Model of Affect has the advantage of treating emotion categories as single points around a circle while at the same time giving sense of location, and ordering, for the emotions. Furthermore, since emotions points are relative to the valence arousal axes, the model gives an easy interpretation of what happens when the valence arousal axes are shifted, or tilted. All this will help us to relate the emotion categories to the *VA* space shortly.

III. Detecting Emotions in Speech

Emotion classification in speech is a challenging task and has received a lot of attention in the past ten years. While there is recent interest in continual modeling of emotions [12], speech utterances are generally assigned to fixed labels, such as Ekman's "big six" emotions (anger, disgust, fear, happiness, sadness and surprise), and emotion speech datasets (corpora) typically contain either acted speech [13], [14] or spontaneous speech [15] assigned to fixed emotion labels.

After any necessary speech-signal pre-processing, low-level feature descriptors are extracted, from which an appropriate model can be constructed. Many parameters are used to detect emotion, including mel-frequency cepstral coefficients (MFCCs), which have been the most investigated features for emotion recognition. MFCCs are simply a compact representation of the spectral envelope of a speech signal. In addition to MFCCs, pitch, intensity, formants and even zero-crossing rate are used. Furthermore the modeling can either be based on fixed length features or variable length features.

Emotions are modeled using a wide variety of techniques including Gaussian mixture models (GMMs), support vector machines and back propagation artificial neural networks. Two recent methods for modeling emotions include class-specific multiple classifiers, based on standard modeling techniques [16], and modeling of emotions using front-end factor analysis (I-vectors) [3], [17].

In the I-vector model, each utterance is expressed as a low-dimensional I-vector (usually between 10 and 300 dimensions). One of the advantages of modeling in the I-vector space is that the I-vectors themselves are generated as unsupervised data [18], without any knowledge of classes. What this essentially means is that when emotion classes are enrolled, a more traditional classifier, such as a Support Vector Machine (SVM) can be used, allowing for quick enrollment of the users' emotional data. This can be an advantage when lots of background data is needed to increase the classification performance. In the I-vector-based system, the background data can be incorporated in the training of the GMM and total variability model, which are used to extract the I-vectors themselves, and which then need not be retrained. Potentially this can reduce modeling of the emotion classifier from hours to seconds. In this work, we have elected to use the I-vector model for emotion classification.

IV. Critique-based Recommender Systems

Since typically the content from only one channel can be consumed at any given point, there is a strong basis for providing recommendation for EPG items by quickly being able to select the most relevant channel.

There have for example been works that have looked at recommending content within the EPG framework, that rely both on collaborative [19] as well as content-based [20] techniques. In particular, collaborative recommender systems rely on using other people's ratings for content to generate a list of recommendations for a user. However, we do not believe these fit in well within the EPG framework. Firstly, there is an out-of-band (with the broadcast stream) exchange of ratings between users that needs to take place. While this may seem trivial with today's permanently connected TVs, it is an overdimensioned solution. Secondly, and most importantly, the very nature of broadcast TV is that much of the content that is broadcast may be short-lived and it is possible that it will never be rebroadcast. Once the program has aired, there would be little interest in other users ratings for the program, had these been collected in the first place.

Knowledge-based recommender systems came into existence to deal with the problem of how to recommend knowledge intensive items such as customizable digital cameras, for which ratings might not be easy to acquire, or where they might not be entirely applicable for the given application [21]. An inherent assumption with knowledge-based systems is that a user may be somewhat undecided on what to search for, and it is therefore the task of the system to guide the user to the item of interest. In a typical case-based recommender, a form of knowledge-based system, a process known as *critiquing* is used in the following manner:

1) The consumer's preference is extracted, either explicitly, or implicitly.
2) Using some sort of similarity metric, the system provides an initial recommendation.
3) The consumer either accepts the recommendation, which ends the entire process, or critiques it, by selecting one of the critique options available.
4) For each critique made, the item space is narrowed down by filtering out the unwanted items, and a new recommendation is made.
5) The process continues until the customer finally selects an item.

A lot of past research has looked at critiquing in the context of high-risk, once off-items, such as digital cameras and automobiles. Since these items are highly customized and often one-off purchases, they require more effort on the part of the user to make a sound decision, since there is a larger penalty to pay if recommendation leads to a poor decision. However, research in a limited capacity has also begun to look at so-called low-involvement product domains [22]. Low-involvement product domains typically entail low-risk items, such as music and TV content. One particular work that is noteworthy in this regard is the MovieTuner feature incorporated into MovieLens, that allows movie qualities, such as *more action* to be adjusted through critiquing [23].

We propose to make use of critiquing to allow navigation of items in the *VA* space, and to gather feedback needed for computing affective offset. By allowing the user themself to take part in the recommendation process gives us feedback on how the user's perceived affective state differs from their desired state, and what they really would like to watch.

V. Recommendation Framework

A. General Overview

A typical system operation can be realized as follows: Once the user's mood has been detected, from audio-based parameters, the closest matching item that matches the user's mood profile is displayed to the user. The user can either accept the item, or request the recommendation of a new item. To be able to make a new recommendation, the user provides information on how the system should constrain its search. The process continues until the user finally accepts the item.

After the recommendation process has completed, the system calculates the affective offset between the initially recommended item and the finally selected item (if any), and takes this into account when processing the output labels from the classification stage, in such a way as to reflect the new mood offset. Fig. 1 shows an overview of the proposed system. We shall now present theory for the individual components.

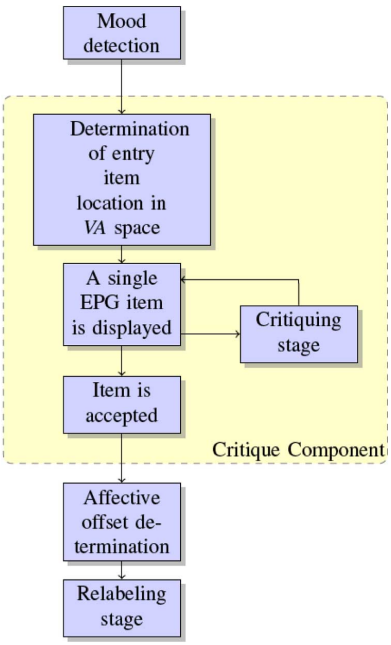


Fig. 1. Complete system overview.

B. Mood Detection

Since it is emotions themselves that are detectable and give rise to moods, we start by discussing emotion detection. Let E be the total number of emotion classes. Emotions can then be detected by analyzing the speech utterances from each user and assigning an emotion class $e \in E$ to each. The more classes that need to be classified, the lower the classification accuracy. What this entails is, that for a set of utterances over a time interval for which the actual emotion was e_a, and the predicted emotion is e_p, there will almost always exist a subset of these utterances where $e_a \neq e_p$, i.e. utterances for which the actual class was not predicted correctly. What is important here is not so much that each emotions is categorized 100% correctly, but that the areas of the emotion space, and hence adjacent emotions, that were detected, are reflected in the profile. With this in mind, the emotion profile for a single user u can be modeled by

$$\vec{e}_u = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_E \end{bmatrix} \tag{1}$$

where p_j, $0 \leq p_j \leq 1$, simply represents the actual predicted probability for emotion class j, $1 \leq j \leq E$, and $\sum_{j=1}^{E} p_j = 1$.

Over a sequence of time intervals, e.g. over the last 12 hours, the system collects the individual emotion profiles, and condenses them to a mood profile

$$\vec{m}_u = \frac{1}{T} \sum_{i=1}^{T} \vec{e}_u i * w_i \tag{2}$$

where

$\vec{e}_u i =$ The $\vec{e}_u$ corresponding to the i time interval
$T =$ Total number of discrete time intervals, and
$w_i =$ Weighting of $\vec{e}_u$ for the i time interval

To compute the weighting, a modified form of the depreciation factor, originally used in computing the depreciation citation count [24] is used to compute w_i.[2] This will ensure that emotions recorded over earlier time intervals, regardless of the size of the time interval, will always contribute less to a given overall mood profile $\vec{m}_u$.

The weighting w_i is thus given by the following:

$$w_i = \frac{1 + \tanh(\frac{i}{T})}{2} \tag{3}$$

C. Determination of Entry Item in Valence Arousal (VA) Space

For a given fixed set of emotions, each emotion can be characterized by associating it with an affective location (offset in degrees) around a circle.[3] There is thus a mapping from each emotion category to its corresponding angle. More formally, this set of emotions can be expressed in the following way:

$$\vec{\Theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_E \end{bmatrix} \tag{4}$$

To map the mood profile $\vec{m}_u$ that was introduced in the previous section, to a point in the *VA* space, we introduce the concept of a *directional mood vector*.

Each component of both $\vec{\Theta}$ and $\vec{m}_u$ is associated with a separate emotion. Therefore for each emotion j, $1 \leq j \leq E$, we create a new vector $\vec{\text{Mood}}_{\text{VA}\,j}$ with magnitude $m_{u\,j}$ and angle θ_j, with the angle measured in degrees from the positive *valence* axis in the *VA* space

$$\vec{\text{Mood}}_{\text{VA}\,j} = [\, m_u j * \cos\theta_j \quad m_u j * \sin\theta_j \,] \tag{5}$$

This results in E separate emotion vectors, where the angle for each serves as an identification for an emotion and the magnitude indicates the confidence of that emotion, as detected by the audio classifier. Finally, all E components are summed to obtain the final directional mood vector. More formally, this is depicted as

$$\vec{\text{Mood}}_{\text{VA}} = \sum_{j=1}^{E} \vec{\text{Mood}}_{\text{VA}\,j} \tag{6}$$

In order to find an appropriate entry item, which forms the first stage of the recommendation process, we associate the directional mood vector with a suitable point in *VA* space. To locate the best item, we iterate through all items, where Γ is the total number of items. For each item $\gamma, \gamma \in \{1, 2, \ldots, \Gamma\}$ a score based on cosine similarity is computed as follows:

$$Score_\gamma = \frac{\vec{Mood}_{\text{VA}} \cdot \vec{k}_\gamma}{\|\vec{Mood}_{\text{VA}}\| \|\vec{k}_\gamma\|} \tag{7}$$

[2]The original depreciation factor is based on years and ours is based on discrete time intervals.

[3]The location of each emotion is determined by past empirical studies [9], as discussed earlier.

where $\vec{k}_\gamma$ is the location of item γ in *VA* space.

The first item to be recommended, or *entry item* is then the item γ which generates the highest score

$$\gamma = \arg\max_{\gamma}(Score_\gamma), \forall\gamma, \gamma \in \{1, 2, \ldots, \Gamma\} \tag{8}$$

D. Critiquing Stage

At this stage the user has the opportunity to examine the entry item.[4] If he/she decides not to accept the item, a critique is specified for the new item. The possible critiques are *more pleasant*, *less pleasant*, *more intense* and *less intense*. These correspond to the affective operations more valence, less valence, more arousal and less arousal, respectively. The algorithm determines beforehand whether there is an availability of items to satisfy the potential constraint. If this condition is not satisfied, the constraint is simply not presented. Although it is possible to implement compound constraints, due to the low dimensionality of the number of free parameters available (only four), we opted for simple constraints only in this work.[5]

Once the user has selected a constraint, the best matching item is determined and displayed in the following way: for a given iteration r, let S be the set of items subject to the new constraint C_r. The next item to be recommended is then the item with the shortest distance between the currently displayed item $item_c$, i.e. the last recommended item, and all other items subject to the constraint, and given as

$$\text{Match} = \min_{\forall s \in S, s \neq c} d(\vec{item}_c, \vec{item}_s) \tag{9}$$

where the distance $d(\vec{item}_c, \vec{item}_s)$ is a weighted form of the standard Euclidean distance in *VA* space

$$d(\vec{item}_c, \vec{item}_i) = \sqrt{w_V * (item_{cv} - item_{iv})^2 + w_A * (item_{ca} - item_{ia})^2} \tag{10}$$

One of the problems with using the standard Euclidean distance is that it is based on pure distance and no consideration is given to the direction in which the user really wishes to traverse the space. Figs. 2 and 3 show the case for a user starting out in the negative valence, positive arousal quadrant (top left), who then executes 14 critique cycles. In every case, the user selects the constraint *more pleasant*, i.e. more valence. For the unweighted case, we note that the user (unintentionally) gradually wanders over to the positive valence / negative arousal quadrant (bottom right), where, ideally, the optimum quadrant would have been the positive valence / positive arousal quadrant (top right). The weights w_V and w_A are therefore introduced and chosen empirically to ensure that more preference is given to either the valence or arousal dimension, depending on what constraint was chosen. This allows for a larger distance in the desired direction to be taken into consideration than would be otherwise, and results in a more direct path. The effect of using these weights is shown in Fig. 3.

[4]The entry item is the very first item that is recommended to the user.

[5]Compound critiques would be suitable if the affective parameters were to be combined with other parameters, such as genre, time of day, and age ratings.

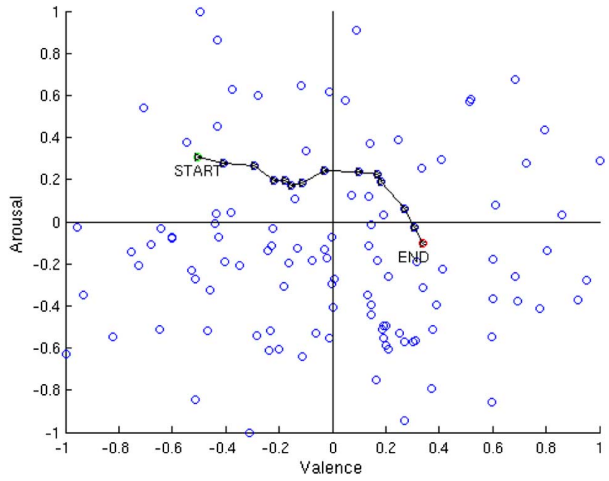


Fig. 2. Navigating the *VA* space before the modified weighted Euclidean distance measure is introduced.

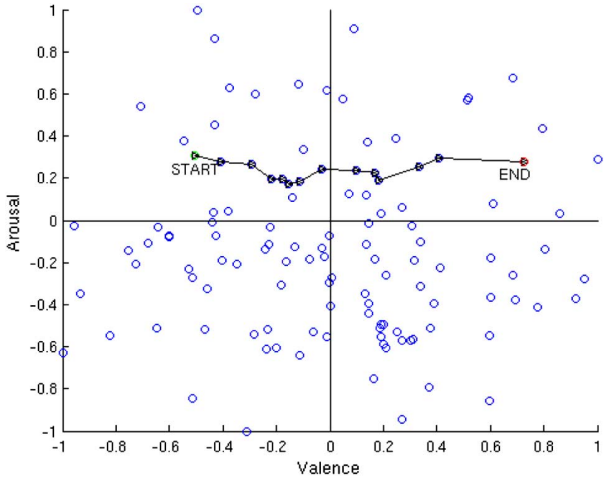


Fig. 3. Navigating the *VA* space after the modified weighted Euclidean distance measure is introduced.

The recommendation process continues until the user selects an item as acceptable, in which case it is terminated.

E. Affective Offset Determination

Once the recommendation process has completed, the user will be located at another point in *VA* space. How far this point is located from the initial recommendation depends on both the number of cycles taken as well as the overall affective bearing the user took. In order to know how far off the user is from the initial recommendation, we now compute the affective offset. This offset will then be taken into account in future recommendation sessions to offset the user's mood profile (the perceived mood) with the recommended content (which relates to the desired mood).

Let A be the vector passing through the origin and the initial point where the user set out from, and let B be the vector passing through the origin and the point representing the finally selected item. The angle of this offset is given as

$$Offset_{Angle} = \arccos\left(\frac{A \cdot B}{\|A\|\|B\|}\right) \tag{11}$$

where

$$\frac{(A \cdot B)}{(\|A\|\|B\|)} = \text{the cosine distance between } A \text{ and } B, \text{ and}$$
$$\arccos(x) = \theta \text{ gives } \theta \text{ in degrees and not radians.}$$

However, not only is the angle important here, but also the direction (on the emotion circumplex) of B relative to A. If we in future recommendation rounds offset the emotions in the wrong direction, instead of compensating for the mismatch between detected mood and recommended item, we would effectively be contributing to the error instead of reducing it.

We therefore determine whether this direction is clockwise, or counter-clockwise. To do this, we first compute the absolute angle of both A and B. The absolute angle for a vector through the origin (positive valence axis) to a given point P, $P = A$, $P = B$, is computed in the following way:

$$Angle_{P_{absolute}} = mod(-\arctan(P_y, P_x) - 90), 360) \quad (12)$$

where

$$\arctan(y, x) = \theta \text{ gives } \theta \text{ indegrees} - (180 \leq \theta \leq 180)$$
$$\text{mod}(\theta, 360) = \theta \text{ gives } \theta \text{ in degrees} (0 \leq \theta \leq 360)$$

Depending on the location of A and B, two possible angles can be computed as

$$Diff_c = mod(Angle_A - Angle_B, 360) \quad (13)$$
$$Diff_{cc} = mod(Angle_B - Angle_A, 360) \quad (14)$$

where

$$Diff_c = B \text{ is located clockwise relative to } A$$
$$Diff_{cc} = B \text{ is located counterc lockwise relative to } A$$

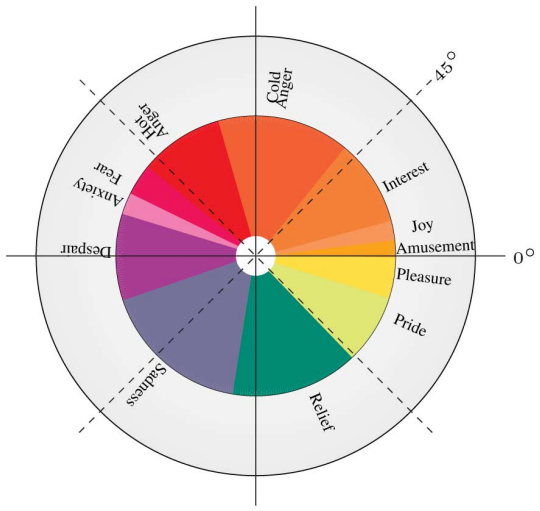
If $Diff_c = Offset_{Angle}$, then this indicates that the offset occurs in the clockwise direction and $Offset_{sign} = -1$. Likewise if $Diff_{cc} = Offset_{Angle}$, then $Offset_{sign} = 1$. The sign is combined with the previously computed offset angle, to give the directional offset

$$Offset = Offset_{Angle} * Offset_{Sign} \quad (15)$$

F. Relabeling Stage

For the 12-class emotion classifier, labels indicate the emotion that each utterance is associated with. There is no concept of distance or overlap between labels - they are simply emotion categories. However, these concepts hold for the emotional spaces themselves, and where they move, so will the labels.

Fig. 4 shows a possible configuration for a set of emotions and their location around the circumplex. For a given configuration, starting from 0°, there exists an explicit fixed ordering of the emotion labels. By tilting the valence and arousal axes by Θ, which happens to be the affective offset calculated in the previous stage, we effectively change the ordering of the labels. An important design consideration was whether to rotate the directional mood vector, as computed in equation 6, or to rotate the labels themselves. The rationale for rotating the speech labels themselves allows for the possibility of incorporating future enrollment data, for example, as might be retrieved through multi-modal emotional systems, and leads to a better accuracy over time. Simply rotating the directional mood vector would make the system unadaptable.



Labels before tilt: Amusement, Joy, Interest,... , Pride, Pleasure
Labels after tilt: Cold Anger, Hot Anger, ..., Joy, Interest

Fig. 4. Tilting of emotion labels. By tilting the valence or arousal axis by Θ, we impose a new ordering of the labels.

Now more formally, let $L = \{l_1, l_2, \ldots, l_E\}$ be the set of labels. Then $X \equiv (l_1, l_2, \ldots, l_E)$ represents the sequence of labels from L before applying the affective offset. The labels in the list are arranged in order of their respective locations starting from $\Theta = \theta_1$. Likewise $Y \equiv (l_1, l_2, \ldots, l_E)$ represents the sequence of labels from L after applying affective offset, but where the list now starts from $\Theta = \theta_2$ instead. The mapping from old label l to new label is then simply carried out by the mapping function $f : L \rightarrow L, l \mapsto Y[Index_X(l)]$, where $Index_X(l)$ is the index of label l in X.

VI. Experimental Work

A. Annotation of Content Items

In an initial user survey, 16 subjects rated 4 sets of 60 TV programs, with 3 subjects being assigned to each set. The TV programs were extracted from the EPG in the interval from 15:00 Friday 13 December 2013 to 10:00 on Saturday 14 December 2013. Each program shown to the evaluator was accompanied by a title, the name of the channel on which it was aired, a two-level category into which the program was placed, for example *"Level1: Movie; Level2: Comedy"*, and finally a short synopsis. All data presented to the evaluators was taken directly from the EPG metadata and was not manipulated by us in any way. The task given for each program was to read the information and thereafter rate the *pervasive mood* of each program in the *VA* space. The method used was the well-known Self Assessment Manikin (SAM) [25], which is a psychological tool used

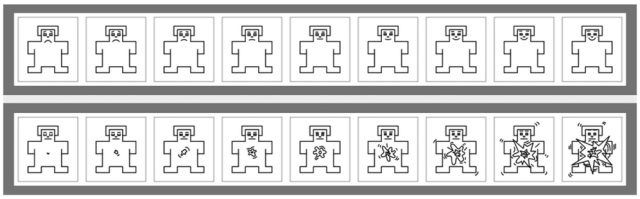


Fig. 5. Scales used to collect the pervasive mood for each TV program. The top scale measures valence and the bottom scale measures arousal. Scales are courtesy of PXLab.[6]

to quantify perceived emotions. It is easy to administer, does not rely on verbs and is suitable for subjects whose native language is not English, which was the case in this study. Subjects were shown a diagram of a 9-point SAM scale, where only valence and arousal ratings were collected. It was possible for subjects to select a point anywhere on the scale, thus allowing collection of continuous valence and arousal values. Subjects were also informed that they would be rating the programs on a continuous scale. Since 3 subjects rated each TV program, a total of three ratings were collected for each. These ratings were averaged, as is customarily done [25], to give a mean SAM ratings for each program. The scales that were used can be seen in Fig. 5.

Once the rating process was complete, the first two sets were combined and the last two sets were combined, yielding two larger sets, A and B of programs containing 120 content items each. The sets were combined in this manner to create a realistically-sized number of items to browse, but taking into account the length of time required to annotate the items.

B. Mood Determination and Audio Classification of Emotions

The audio data used to represent the home user's emotional state was taken from the Geneva Multimodal Emotion Portrayals (GEMEP) [14], which was also chosen as the dataset for the emotion sub-challenge part of the Interspeech 2013 Computational Paralinguistics Challenge [27]. The dataset contains 1260 short voice utterances, divided into 18 emotional classes. The data is split across 10 actors, of which half are male and other half female. Due to the fact that 6 out of 18 of the emotions occur very sparsely in the dataset, the classification was restricted to 12 separate emotions. These were amusement, pride, joy and interest (positive valence, positive arousal), anger, fear, irritation and anxiety (negative valence, positive arousal), despair and sadness (negative valence, negative arousal) and finally pleasure and relief (positive valence, negative arousal). One of the primary reasons for selecting the GEMEP corpus was its wide spectrum of available emotions.

For each case, we connected the mood configuration to real speech utterances from the dataset by assigning each mood to the most appropriate emotions. The good mood was associated with the emotions *amusement*, *joy* and *interest*, the bad mood was associated with *cold anger (irritation)*, *hot anger*, *fear*, *despair*, *anxiety* and *sadness*, and the neutral mood was associated with the emotions *relief*, *pride* and *pleasure*. For each test trial, a speaker was randomly identified from the GEMEP dataset and a mood configuration was selected. The relevant emotion features, taken from the test set, were then concatenated and used for mood profile determination.

12-way classification of the data was carried out using front-end factor analysis (I-vectors), using the ALIZE 3.0 framework [28]. The process was as follows: 13 MFCCs (including log energy), first and second derivatives were extracted to give a fixed 39-feature frame for each 25 ms voice frame, with a 10 ms overlap for each frame. A 128-component Gaussian mixture model (GMM) was trained with the entire training set. At this point, the six unused classes were not utilized further in the system. Using the data from the GMM, a total variability matrix was trained. Subsequent to this, for each utterance, a 90-dimensional I-Vector was extracted from the total variability matrix. Once in the I-vector space, classification of the utterances was then carried out using probabilistic linear discriminant analysis (PLDA) after performing normalization on the I-vectors. PLDA is known to exhibit good performance when used for the classification of I-Vectors. The accuracy for the acoustic sub-system for all 12 classes on the development set was 42.72%, and on the test set (used in the end-to-end system) was 41.20%, which is in line with the state-of-the-art [27], [29]. More detailed results for the individual categories can be seen in Table I.

C. Other System Parameters

The affective state locations used for computing the directional mood vector were adopted from past studies [9]. The range of values and calculated mean for each emotion is shown in Table II. For the operations *more pleasant* and *less pleasant* the weights were set to $w_v = 0.45$ and $w_a = 1$ and for the operations *more intensity* and *less intensity* the weights were set to $w_v = 1$ and $w_a = 0.45$. The user interface used for presenting the items to users and used for critiquing was implemented in PHP.[7]

D. User Evaluation

26 subjects agreed to take part in a series of six evaluations, with each evaluation carried out on a separate day. Half of the subjects were assigned the items from group A and the other half were assigned the items from group B[7]. Each person was given access to a web portal through which they could interact with the critique-based recommender.

Each evaluation considered a single mood case. On the first day, each subject was told to picture themselves being in a neutral and relaxed mood, and to strengthen their mood, they were presented with a set of 10 neutrally rated pictures, taken from the International Affective Picture System (IAPS) [30]. The IAPS is a database of colour photographs, each annotated with valence, arousal and dominance ratings, and which is often used to elicit emotions in affective-related research studies. The subject was asked to not spend more than 15 seconds viewing each picture.

Once all pictures had been viewed, the information for a TV program was then presented to the user. The subject was asked to rate the program, on a scale of 1 to 10, on how suitable they thought the program was for the given mood.

[6]The PXLab Self Assessment Manikin Scales. Available: http://irtel.uni-mannheim.de/pxlab/demos/index_SAM.html

[7]PHP stands for "PHP: Hypertext Processor." It is a server-side scripting language suitable for the implementation of interactive web deployments.

TABLE I
CONFUSION MATRIX FOR 12-CLASS EMOTION CLASSIFIER. SHADED ENTRIES CORRESPOND TO ACTUAL class = predicted class

	Amusement	Anxiety	Irritation	Desperation	Joy	Anger	Interest	Fear	Pleasure	Pride	Relief	Sadness
Amusement	90.00	0.00	0.00	3.33	0.00	0.00	0.00	6.67	0.00	0.00	0.00	0.00
Anxiety	13.33	23.33	3.33	6.67	3.33	6.67	3.33	10.00	6.67	10.00	0.00	13.33
Irritation	3.12	6.25	18.75	6.25	0.00	0.00	9.38	6.25	0.00	15.62	15.62	18.75
Desperation	33.33	10.00	3.33	23.33	0.00	3.33	0.00	16.67	0.00	3.33	0.00	6.67
Joy	20.00	0.00	3.33	6.67	23.33	13.33	0.00	20.00	0.00	3.33	0.00	10.00
Anger	6.67	6.67	3.33	3.33	3.33	56.67	6.67	6.67	0.00	3.33	0.00	3.33
Interest	0.00	3.33	0.00	0.00	0.00	0.00	16.67	3.33	20.00	0.00	10.00	46.67
Fear	3.33	6.67	0.00	10.00	6.67	13.33	0.00	46.67	0.00	6.67	6.67	0.00
Pleasure	3.33	3.33	3.33	3.33	0.00	0.00	13.33	0.00	43.33	0.00	6.67	23.33
Pride	16.67	0.00	6.67	6.67	0.00	13.33	0.00	0.00	3.33	40.00	3.33	10.00
Relief	0.00	10.00	0.00	6.67	0.00	3.33	10.00	0.00	23.33	3.33	43.33	0.00
Sadness	3.45	6.90	3.45	3.45	0.00	0.00	0.00	0.00	3.45	6.90	3.45	68.97

TABLE II
AFFECTIVE STATE LOCATIONS [9]

Emotion	Range	Mean Value
Amusement (delight)	6-12	7
Joy	5-10	7.5
Interest (activation)	20-36	28
Cold anger (irritation)	88	88
Hot anger	83-171	127
Fear	141-161	151
Anxiety (worry)	149-163	156
Despair (discouragement)	163-173	168
Sadness	144-311	227
Relief (relaxation)	249-338	293
Pride	309-3	336
Pleasure (Contentment)	347-359	353

After the program had been rated, it could either be accepted, in which case the recommendation session for that program was over, or the user could select a better recommendation case (critique the current item). If they chose to critique the item, they were then presented with a list of choices for selecting a more pleasant, less pleasant, more intense or less intense item. For each option the number of items available for that selection were also displayed, giving the evaluator an updated indication of the potential items in each direction. In the case where no items were available, the option to navigate in that direction was not presented to the user. Furthermore, users were not prevented from navigating back over the same items they had seen before. Each new item that was presented constituted a new critiquing iteration. The number of iterations it took the user to finally select an item (by marking "accept") was counted and stored. Finally, for the final item, the user was asked once again to rate the item on a scale of 1-10, on how suitable they thought the item was.

After the subject had completed the first part of the evaluation (the neutral mood case), they were allowed to move onto the next part. Days 2 and 3 were identical to Day 1, except that different mood settings were used. For Day 2, the subject was told to instead imagine being in a good mood, where correspondingly good mood pictures were shown. In a similar fashion, Day 3 followed where the subject was now told to imagine being in a bad mood, with correspondingly bad mood pictures being shown.[8] Furthermore subjects were not allowed to complete two parts on the same day. If a subject skipped a day, a follow-up mail was sent to them. After two follow-up mails had been sent, with no response, the subject was considered to have abandoned the survey. Four of the subjects ended up dropping out of the survey, and hence we only present data for 22 out of the initial 26.

For days 4, 5 and 6, subjects were asked to repeat the evaluation for the neutral, good and bad mood cases, respectively. However, on these days, the users were not informed that their affective offset from the previous round (matching that particular mood), had been recorded and used to offset the system. Users were shown a new set of frame slides for the second round of each mood case, since seeing the same slides again would have a reduced effect.

[8]The following IAPS pictures were used in the evaluations: Days 1, 2, and 3, Neutral Mood: 1121, 1616, 2102, 2221, 2377, 2575, 2579, 2745.1, 7497, 7503; Good Mood: 2216, 2598, 4614, 5210, 5814, 7405, 7508, 8034, 8503, 8531; Bad Mood: 2205, 2456, 2745.2, 2751, 6313, 9185, 9252, 9635.1, 9904, 9940 - Days 4, 5, and 6, Neutral Mood: 2026, 2036, 2377, 2382, 2383, 2410, 2594, 2840, 7003, 7640; Good Mood: 1722, 1811, 2158, 7492, 8090, 8163, 8300, 8350, 8420, 8510; Bad Mood: 2399, 2682, 2683, 2703, 2800, 2900, 6021, 9420, 6570.1, 9908.

VII. RESULTS AND DISCUSSION

A. Effect on the Number of Iterations

In Table III, we show a summary of results for both evaluations (before and after applying affective offset) and for all three mood cases. When browsing in the *VA* space to find more suitable items, users can revisit older items as many times as they wish (in case they change their mind). Occasionally, this leads to the path from initial item to final item containing one or several loops, e.g. if while browsing, a user visits items $\{A, B, C, D, E, C\}$, the loop $\{C, D, E, C\}$ can be replaced with $\{C\}$ giving the shorter path $\{A, B, C\}$. For the sake of brevity, we refer to paths including loops as *full paths* and paths with the loops removed as *direct paths*. We are interested in these direct paths since going around in a loop essentially means the user ended up at the same spot they were at previously, and hence the same region. Direct paths are therefore a better summary of a user's ultimate migration. We therefore show results for both types of path, where the first two rows in Table III show the number of iterations for full paths, and the next two rows show results for the number of iterations for direct paths. The following two rows then show the ratio between direct path and full path - the closer to 1 the ratio is, the fewer the number of loops, and the more direct the full path is. The final row shows the average affective offset for each case (ignoring the direction of the offset).

Looking at the number of iterations on average that were taken to find a suitable item, in all cases, as shown in Table III, we can see that a lower number of iterations was needed in the case where the user's affective offset was applied. An overall improvement was obtained of 43.60% and for the good

TABLE III
SUMMARY OF RESULTS FOR BOTH EVALUATIONS (BEFORE AND AFTER APPLYING AFFECTIVE OFFSET) AND FOR ALL THREE MOOD CASES. M = Mean, SD = Standard Deviation

	Good (M)	Good (SD)	Bad (M)	Bad (SD)	Neutral (M)	Neutral (SD)	Overall (M)	Overall (SD)
Avg iterations before (full path)	19.55	15.70	34.55	43.49	21.18	20.48	25.09	29.53
Avg iterations after (full path)	9.36	8.58	14.68	13.21	18.41	18.70	14.15	14.39
Avg iterations before (direct path)	9.64	5.49	10.36	8.44	8.82	10.29	9.61	8.21
Avg iterations after (direct path)	5.68	4.03	6.14	4.23	8.09	7.81	6.64	5.64
Ratio (direct path / full path) before	0.49	0.35	0.30	0.19	0.42	0.50	0.38	0.28
Ratio (direct path / full path) after	0.61	0.47	0.42	0.32	0.44	0.42	0.47	0.39
Affective offset (ignoring direction)	68.96	57.33	49.71	41.65	42.09	47.05	53.59	49.67

mood, bad mood and neutral mood cases, improvements were obtained of 52.12%, 57.51% and 13.08% respectively. The average number of iterations when applying affective offset was significantly lower than when it was not applied[9] ($z = -3.39$, $p < 0.01$, $r = -0.51$). The reduction in iterations was also significant for the both the good mood case ($z = -2.90$, $p < 0.01$, $r = -0.44$), and the bad mood case ($z = -2.40$, $p < 0.05$, $r = -0.36$). However, for the neutral mood case, it was not significant ($z = -0.80$, $p = 0.42$, $r = -0.12$).

We believe the initial rather large standard deviation is due to the fact that browsing is rather personal. Some people generally tend to browse more than others. If browsing is indeed personal, then a Pearson correlation between the before and after iterations for each mood and all users should reveal a medium to large effect size. Conducting such a correlation gives values of $r = 0.28$ for the overall case, $r = 0.44$ for the good mood case, $r = 0.38$ for the bad mood case and $r = 0.18$ for the neutral mood case. The fairly strong relationship for the good and bad mood cases indicates that users are definitely more consistent in their behavior in these mood cases than in the neutral case (and more so in the good mood case).

For the direct paths, we find an overall improvement of 30.91%, and improvements of 41.08%, 40.73% and 8.28%, for the good, bad and neutral mood cases, respectively. Once again, the overall reduction was significant ($z = -2.61, p < 0.01, r = -0.39$). It was also significant for the good mood case ($z = -2.43, p < 0.05, r = -0.37$), the bad case ($z = -2.13, p < 0.5, r = -0.32$), but not significant for the neutral mood case ($z = -0.09, p = 0.93, r = -0.01$). These results indicate that even in the absence of loops, there is still a significant reduction in the path length. The higher direct path / full path ratios, for all mood cases, after applying affective offset, indicates fewer loops and more direct browsing paths.

Looking at the affective offset that arose in each case, we see the exact same trend as was seen for both full paths, direct paths, and user consistency, in terms of their statistical power. The largest affective offset was 69.96 degrees for the good mood case, followed by 49.71 degrees for the bad mood case, and finally 42.09 degrees for the neutral mood case.

These results are interesting when seen in light of the free-style user feedback comments that some of the participants provided. Four people wrote that they found it difficult to place themselves in a neutral mood setting and that the good mood setting was far easier to relate to. This might explain why in the neutral mood case there was no significant reduction in iterations - the confusing neutral mood setting resulted in participants being less consistent than in the other mood cases. The bad mood case was also considered easier to relate to, but people had more to say in general on what they thought was appropriate content for this mood. Three participants said they would only consider content that would repair their bad mood state, two wrote that comedies would be ideal, one person wrote that more intense content would be a good choice, and another two reported that if they were in a bad mood, they would not watch TV at all. It seems that the bad mood case is possibly less natural than the good mood cause and causes people to think more about what they want to watch. In the good mood case, people seem to be more open as what they want to see, and suggesting a good region allows them to more quickly find an item. In the bad mood case however, people are fussier about what they want to see - even when the region is right, more browsing is needed to find a good item.

B. *Effect on User Ratings*

In both evaluations, and for all mood cases, users were asked to rate both the initially recommended item as well as the finally selected item on a scale of 1 to 10, on how good a match they though the items were. A summary of the results for these ratings is shown in Table IV. The first two rows cover the first evaluation before affective offset and the second two rows cover the second evaluation after affective offset.

Firstly, as expected, the final items for each evaluation were rated higher than the initial items, and in all cases these were significant: For the first evaluation, the overall increase went from 5.07 to 7.80 ($z = -5.86$, $p < 0.1$, $r = -0.88$), for the good mood case 4.95 to 7.91 ($z = -3.48$, $p < 0.01$, $r = -0.52$), for the bad mood case from 4.73 to 7.32 ($z = -3.24$, $p < 0.01$, $r = -0.49$) and for the neutral mood case 5.55 to 8.18 ($z = -3.48$, $p < 0.01$, $r = -0.53$. Likewise for the

[9]Treating the null hypothesis that the difference between the number of iterations before and after comes from a distribution of zero median, we use the sign rank test to test for significance. The effect size is computed as $r = Z/\sqrt{N}$ (Z is the Z-score and N is the observation count (22 users gives 44 observations)). The interpretation of r goes according to Cohen's benchmark (where a potential minus sign is ignored): $r > 0.1$ is a small effect size, $r > 0.3$ is a medium effect size and $r > 0.5$ is a large effect size.

TABLE IV
SUMMARY OF RATINGS FOR BOTH EVALUATIONS AND FOR ALL THREE MOOD CASES. $\mathrm{M} = \mathrm{Mean}, \mathrm{SD} = \mathrm{Standard\ Deviation}$

	Good (M)	Good (SD)	Bad (M)	Bad (SD)	Neutral (M)	Neutral (SD)	Overall (M)	Overall (SD)
Avg rating, Initial item, 1st evaluation	4.95	2.30	4.73	2.39	5.55	2.65	5.07	2.44
Avg rating, Final item, 1st evaluation	7.91	1.51	7.32	1.81	8.18	1.84	7.80	1.74
Avg rating, Initial item, 2nd evaluation	5.23	3.18	5.64	2.63	6.91	2.51	5.92	2.84
Avg rating, Final item, 2nd evaluation	8.91	0.92	8.32	1.17	8.73	1.12	8.65	1.09

second evaluation, the overall increase went from 5.92 to 8.65 ($z = -5.82$, $p < 0.01$, $r = -0.88$), for the good mood case, from 5.23 to to 8.91 ($z = -3.44$, $p < 0.01$, $r = -0.52$), for the bad mood case from 5.64 to 8.32 ($z = -3.46$, $p < 0.1$, $r = -0.52$) and for the neutral mood case 6.91 to 8.73 ($z = -3.30$, $p < 0.01$, $r = -0.50$). This indicates that browsing was effective enough to find more suitable items.

We also looked at the ratings for only the initial item for both evaluation rounds, and found that in all mood cases, that the initial item in the second evaluation round received a higher rating that in the first evaluation round. However, in none of the mood cases was this actually significant.

More interesting though are the final ratings for both evaluation rounds. Here we found that the final ratings, after browsing had taken place, increased overall from 7.80 to 8.65 ($z = -3.54$, $p < 0.01$, $r = -0.53$), for the good mood case from 7.91 to 8.91 ($z = -2.34, p < 0.5, r = -0.35$), for the bad mood case from 7.32 to 8.82 ($z = -2.41, p < 0.5, r = -0.36$), and for the neutral mood case from 8.18 to 8.73 ($z = -1.40$, $p = 0.14$, $r = -0.21$), which were not significant. The good and bad ratings being strongly significant, and the neutral ratings not being significant suggests a link between ratings and reduction of iterations - in the neutral case users took longer to find an item they really liked (or they simply gave up), which in turn explains the low iteration reduction. The lower standard deviation for all mood cases, as noted by comparing the final ratings for both evaluations, suggests more user consensus in the higher ratings for the second evaluation than in the first. The combination of affective offset and browsing might have a stabilizing effect on users' rating behavior. We emphasize furthermore that users were not shown their previous ratings at all, and since the evaluations were carried out on separate days, would have been unlikely to recall their previous ratings. Nevertheless, in all cases we note that the final average ratings for the second evaluation were higher than *any* of the other three ratings, indicating the point of ultimate satisfaction.

The less significant initial ratings imply that applying affective offset does not necessarily help to improve the *initially* recommended item, but given the added browsing functionality, allows a good *final* item to be located. This is an interesting finding because it indicates that single-shot recommendation of items based on users' audio features is not quite adequate. For example, a user in a good mood might be recommended an emotionally appropriate item, such as a sports game. However, if they are not interested in sports, regardless of the accuracy of the match, the item is likely to receive a low rating. It therefore makes sense to rather recommend a *region* from which the search is to be commenced, and then to harness the particular user's feedback to provide a better (more personal) recommendation the next time round.

C. Effect of Audio Classification

To show qualitatively how our system is affected by the inaccuracies of the audio classification component, we briefly turn our attention to six examples that show how the directional mood vector $\vec{Mood}_{\mathrm{VA}}$ (Equation (6)), changes with label rotations, all of which can be seen in Table V. To recap, a change in the configuration of emotion labels leads to a different placement of the directional mood vector, and hence determines the initially recommended item. A value of 0 indicates no rotations and corresponds to the label sequence 'amu-joi-int-irr-col-peu-inq-des-tri-sou-fie-pla'. This corresponds to observing a very low affective offset or when the finally selected item remained in the same emotion region. A value of 1 indicates one displacement and the sequence 'pla-amu-joi-int-irr-col-peu-inq-des-tri-sou-fie', a value of 2 the sequence 'fie-pla-amu-joi-int-irr-col-peu-inq-des-tri-sou' and so on. If the offset is in the other direction, the label shifting is reversed. From the figures three things are apparent:

1) Shifting of labels does not necessarily lead to an even displacement of the directional the mood vectors around the circle. This is particularly evident for the bad mood case for the female speaker.
2) Occasionally the ordering of labels is not preserved. This can be seen in the bad mood case for the female speaker and in the neutral mood case for the male speaker.
3) In some cases certain areas of the emotion space appear to be underrepresented. This can be seen in the bad mood case for the female speaker, where a large potential area for content items might be excluded.

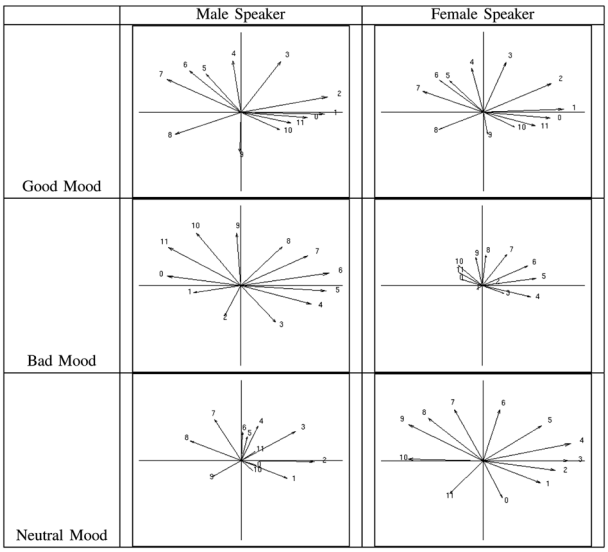
The primary cause for these effects is due to the limited performance of the audio classifier. Since each test trial contains multiple speaker utterances, the limited accuracy of the emotion classifier causes utterances to fall into different emotion categories, which then contribute to unwanted shifting of the directional mood vector. Furthermore the emotion coordinates given in Table II are not evenly spaced apart, which further contributes to the above-mentioned effects.

D. Limitations of the Model and Our Study

Finally, we observed five issues with the proposed model and experiments that we think are worthy of discussion:

1) The model does not handle items situated close to the VA origin very well. Take for example the case where the currently selected item is located in the positive valence,

TABLE V
EXAMPLES OF THE DIRECTIONAL MOOD VECTOR FOR 12 DIFFERENT LABEL DISPLACEMENTS FOR THE THREE MOOD CASES



positive arousal quadrant, and where just a few browse operations leads the user to the negative valence, negative arousal quadrant. Although the distance between these items may be quite short, the resulting affective offset might be quite large. Another problem with this model is that rotation of labels will only occur when a user has moved far enough to wander into a new emotion region. For the proposed emotion offsets given in Table II, some areas are larger than others, meaning that more browsing will be needed to trigger a rotation.

2) As also seen in both Table I and Table V, the effectiveness of the model is affected by the accuracy with which individual emotions can be recognized.
3) The three mood profiles for each user are assumed to be fixed. However, it is possible that some users' mood profiles might vary over time.
4) One of the problems faced with the user evaluation itself is that three of the subjects wrote that they found it difficult to browse programs in the neutral mood setting, and that it was far easier to imagine a good or bad mood case. As evidenced by the results, this difficulty in relating to the neutral mood setting almost certainly led to the rather poor results across the board for the neutral mood setting. It appears that users perform better in a more activated mood state.
5) Two people complained that they did not necessarily always agree with the valence and intensity of programs that the initial subjects had rated, indicating just how personal each user's taste is, and also raises the question of the effectiveness of using third party annotations.

VIII. CONCLUSION

In this paper we developed a framework for recommending TV content based on moods derived from user's emotions. By allowing the user to take part in the recommendation process, we were able to compute each user's affective offset, to be used for future recommendation sessions. We used each user's affective offset to locate an initial region for recommendation, from which a recommendation was determined. The use of affective offset led to better user satisfaction overall, where ratings went from 7.80 up to 8.65. Furthermore, there was a marked decrease in the number of cycles that was needed to find a good item, compared to the case when no affective offset was applied, which went from 29.53 down to 14.39. Future work could include better modeling of items situated close to the VA origin, more predictive modeling of the directional mood vector and a framework that takes in account mood profiles that vary over time.

ACKNOWLEDGMENT

The authors would like to thank the Swiss Center for Affective Sciences for allowing us to use the GEMEP database. We would also like to thank Gracenotes for providing the relevant tools with which to extract the EPG data.

REFERENCES

[1] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.
[2] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151–167, 2013.
[3] R. Xia and Y. Liu, "Using I-vector space model for emotion recognition," in *Proc. Interspeech*, 2012, pp. 2230–2233.
[4] M. Tkalcic, A. Kosir, and J. Tasic, "Affective recommender systems: The role of emotions in recommender systems," in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 9–13.
[5] P. Ekman, "Moods, emotions, and traits," in *The Nature of Emotion: Fundamental Questions*, P. Ekman and R. Davidson, Eds. New York, NY, USA: Oxford Univ. Press, 1994, pp. 56–58.
[6] P. Winoto and T. Y. Tang, "The role of user mood in movie recommendations," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 6086–6092, 2010.
[7] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," *Interact. Comput.*, vol. 18, no. 2, pp. 139–170, 2006.
[8] J. A. Russel, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1170, 1980.
[9] N. A. Remington, L. R. Fabrigar, and P. S. Visser, "Reexamining the circumplex model of affect," *J. Personality Social Psychol.*, vol. 79, no. 2, p. 286, 2000.
[10] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *J. Personality Social Psychol.*, vol. 76, no. 5, p. 805, 1999.
[11] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Proc. 2009 IEEE Int. Conf. Multimedia Expo*, 2009, pp. 566–569.
[12] B. Schuller *et al.*, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4–39, 2013.
[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.
[14] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, p. 1161, 2012.
[15] A. Batliner *et al.*, ""You stupid tin box" – Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. LREC*, 2004, pp. 171–174.
[16] A. Milton and S. Tamil Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," in *Comput. Speech Lang.*, 2013.
[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans., Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[18] D. Martnez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 861–864.
[19] A. B. Barragáns-Martínez, E. Costa-Montenegro, J. C. Burguillo, M. Rey-López, F. A. Mikic-Fonte, and A. Peleteiro, "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition," *Inf. Sci.*, vol. 180, no. 22, pp. 4290–4311, 2010.
[20] M. Bjelica, "Unobtrusive relevance feedback for personalized TV program guides," *IEEE Trans. Consumer Electron.*, vol. 57, no. 2, pp. 658–663, 2011.
[21] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[22] L. Chen and P. Pu, "Critiquing-based recommenders: Survey and emerging trends," *User Model. User-Adapt. Interact.*, vol. 22, no. 1–2, pp. 125–150, 2012.
[23] J. Vig, S. Sen, and J. Riedl, "Navigating the tag genome," in *Proc. 16th Int. Conf. Intell. User Interfaces*, 2011, pp. 93–102.
[24] E. Amolochitis, I. T. Christou, Z.-H. Tan, and R. Prasad, "A heuristic hierarchical scheme for academic search and retrieval," *Inf. Process. Manage.*, vol. 49, no. 6, pp. 1326–1343, 2013.
[25] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Exp. Psych.*, vol. 25, no. 1, pp. 49–59, 1994.
[26] H. Irtel, *The PXLab Self Assessment Manikin Scales*, 2008 [Online]. Available: http://irtel.uni-mannheim.de/pxlab/demos/index_SAM.html
[27] B. Schuller *et al.*, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech 2013*, 2013, pp. 148–152.
[28] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. Mason, and J.-Y. Parfait, ValidSoft UK Ltd., London, U.K., "ALIZE 3.0—open source toolkit for state-of-the-art speaker recognition," in *Proc. Interspeech 2013*, 2013, pp. 1–5.
[29] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using AdaBoost," in *Proc. Interspeech 2013*, 2013, pp. 1–5.
[30] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Instruction manual and affective ratings," Center Res. Psychophysiol., Univ. Florida, Gainesville, FL, USA, Tech. Rep. A-4, 1999.

Sven Ewan Shepstone (M'11) received the B.S. and M.S. degrees in electrical engineering from the University of Cape Town, Cape Town, South Africa, in 1999 and 2002, respectively. He is an industrial Ph.D. candidate at Aalborg University, Aalborg, Denmark.

He is currently with Bang and Olufsen A/S, Struer, Denmark. His main research interests are digital TV and the application of speech technologies to recommender systems.

Zheng-Hua Tan (M'00–SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999.

Previously, he was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, an Associate Professor in the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, and a Postdoctoral Fellow in the Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea. Since May 2001, he has been an Associate Professor in the Department of Electronic Systems, Aalborg University, Aalborg, Denmark. His research interests include speech and speaker recognition, noise robust speech processing, multimedia signal and information processing, multimodal human-computer interaction, and machine learning.

Dr. Tan has published extensively in the aforementioned areas in refereed journals and conference proceedings. He is an Editorial Board Member/Associate Editor for *Elsevier Computer Speech and Language*, *Elsevier Digital Signal Processing*, and *Elsevier Computers and Electrical Engineering*. He was a Lead Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He has served as a program co-chair, area and session chair, tutorial speaker, and committee member in many major international conferences.

Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988, and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, Denmark, in 1995.

He was previously with the Telecommunications Laboratory, Telecom Denmark, Ltd., Copenhagen, Denmark; the Electronics Institute, Technical University of Denmark, Lyngby, Denmark; the Scientific Computing Group, Danish Computing Center for Research and Education, Lyngby, Denmark; the Electrical Engineering Department, Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for Person Kommunikation (CPK), Aalborg University, Aalborg, Denmark. He is currently a Full Professor with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, and is heading a research team working in the areas of numerical algorithms, optimization, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.

Prof. Jensen was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and *Elsevier Signal Processing*, and is currently Member of the Editorial Board of *EURASIP Journal on Advances in Signal Processing*. He is a recipient of a European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section, and Founder and Chairman of the IEEE Denmark Sections Signal Processing Chapter. In January 2011 he was appointed as a member of the Danish Council for Independent Research - Technology and Production Sciences by the Danish Minister for Science, Technology, and Innovation.