Audio-based Age and Gender Identification to Enhance the Recommendation of TV Content

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

Abstract — *Recommending TV content to groups of viewers is best carried out when relevant information such as the demographics of the group is available. However, it can be difficult and time consuming to extract information for every user in the group. This paper shows how an audio analysis of the age and gender of a group of users watching the TV can be used for recommending a sequence of N short TV content items for the group. First, a state of the art audio-based classifier determines the age and gender of each user in an M-user group and creates a group profile. A genetic recommender algorithm then selects for each user in the profile, a single personalized multimedia item for viewing. When the number of items to be presented is different to the number of viewers in the group, i.e. M ≠ N, a novel adaptation algorithm is proposed that first converts the M-user group profile to an N-slot content profile, thus ensuring that items are proportionally allocated to users with respect to their demographic categorization. The proposed system is compared to an ideal system where the group demographics are provided explicitly. Results using real speaker utterances show that, in spite of the inaccuracies of state-of-the-art age-and-gender detection systems, the proposed system has a significant ability to predict an item with a matching age and gender category. User studies were conducted where subjects were asked to rate a sequence of advertisements, where half of the advertisements were randomly selected, and the other half were selected using the audio-derived demographics. The recommended advertisements received a significant higher median rating of 7.75, as opposed to 4.25 for the randomly selected advertisements*[1].

Index Terms — demographic filtering, genetic algorithms, age identification, gender identification, proportional recommendation, advertisement

I. Introduction

With the merging of DVB-C, DVB-T and DVB-S technologies in recent TV platforms, consumers have become overwhelmed by the sheer amount of content available. A large body of research has therefore looked at various ways of personalizing TV to match the needs of users as far as possible. The Electronic Program Guide (EPG), now an integrated component of most modern television sets, has helped to a small extent to narrow the selection of upcoming programs, and there have been various attempts to personalize the EPG to make it even more effective. However, personalization is not only relevant for the EPG, which is geared towards displaying items that can be selected or scheduled, but also for more dynamic content, such as advertising, trailers and short news clips, which are the glue between program segments on broadcast TV.

In order to be able to recommend content, a user profile is needed. User profiles for recommendation can be extracted explicitly, e.g. through registration questionnaires [1] or by asking users to provide ratings. Data can also be collected implicitly through usage patterns [2], [3], and subsequently fed to a central information server, which can then make recommendations.

Traditional recommender systems then use these profiles, together with meta-data and ratings from other users in the network, to provide personalization. One of the issues however, in the context of broadcast TV, is the lack of an uplink channel, through which information such as ratings can be exchanged with the remaining users. It is therefore highly desirable that feedback from users be collected locally, in the set-top box or smart TV if possible, and as unobtrusively as possible, e.g. such as through unobtrusive relevance feedback [3].

By means of local recommendation and implicit user feedback, these systems can work quite effectively, but it is important to consider the preferences of a group of users as well as a single user. This is a particular issue when multiple consumers share a single device, such as a home television, but each has their own user profile and tastes [4]. In the Socially Aware TV Program Recommender for example [5], groups of users who want simultaneous access to the TV are taken into account, where individual profiles that have a common interest are combined.

What is more challenging however, is when multiple viewers share the same TV, but typically only use one person's login, even when a multiple login feature exists, making specification of demographics, extraction of ratings or monitoring for each user difficult to realize in practice. Groups of viewers are further characterized by the fact that users

[1] This work was supported by the Danish Ministry of Science, Innovation and Higher Education under Grant no 12-122802, and by Bang and Olufsen A/S

Sven Ewan Shepstone is with Bang and Olufsen A/S, Peter Bangs Vej 15, 7600 Struer, Denmark (e-mail: ssh@bang-olufsen.dk).

Zheng-Hua Tan is with the Department of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark (e-mail: zt@es.aau.dk).

Søren Holdt Jensen is with the Department of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark (e-mail: zt@es.aau.dk).

Contributed Paper
Original manuscript received 06/22/13
Revised manuscript received 06/30/13
Current version published 09/25/13
Electronic version published 09/25/13.

continually come and go, meaning that the TV must quickly adapt itself to the current configuration.

Taking these multiple requirements into account, i.e. local recommendations, implicit gathering of user information and being able to support groups of viewers, one area that has been somewhat overlooked in the context of personalization of multimedia content, is the home audio environment, from which a wealth of user information can be extracted. State-of-the-art feature extraction and modeling techniques, which are in many ways similar to speaker identification systems, make it possible to extract a number of useful attributes from home viewers, from which recommendation profiles can be constructed. In particular, both the age and gender of TV viewers can be extracted.

Determining both the age and gender of speakers is a complicated task and has received considerable attention in recent years. The results achieved are encouraging and are beginning to make it feasible to use this technology as a viable alternative to existing methods of providing user demographics. Age and gender classification systems are generally implemented as a fusion of several subsystems [6], with each subsystem operating using a form of Gaussian mixture model, multilayer perceptrons, hidden Markov models and/or support vector machines [7], [8].

Recent work shows that 3-class gender detection can be done with an accuracy as high as 75 %, which is roughly 30 % higher than results achieved for 4-class age detection [6]. Here, results for a 7-class classification system also show that separate classes defined as children, young males, young females, adult males, adult females, senior males and senior females can be detected with 61.0 %, 49.4 %, 57.1 %, 27.1 %, 33.8 %, 69.7 % and 53.9 % accuracy, respectively. The largest confusion occurs between young males and adult males, between young females and adult females, between adult female and senior females and finally between children and young females. Even though a lot of room remains for future research to improve these results, there ought to be a substantial basis for recommendation, since the effect of overlapping confusion classes could well be ameliorated by soft market boundaries. For example, in an advertising context, there are many products that would be recommended to both young females and adult females, thus helping to cancel out the confusion overlap seen in these results.

The contribution of this paper is a novel method using on-the-fly detection of the age and gender of the audience present to quickly provide recommendations of TV content to home viewer groups. This is in contrast to other methods that make use of usage data, registration data or questionnaires to obtain the demographics. The focus is on groups of users who are about to be presented with a series of short media items, e.g. between programs. In particular, this work will focus on recommending sequences of advertisements to viewers. The proposed system operates by determining the age and gender class of each user in the group, and then uses this information to find a sequence of content items that best matches the group profile. Ideally, each user should be matched with a content item that belongs to the same age and gender category as that of the user themself. Since the number of advertisements is often predetermined in advance and may not be equal to the number of viewers present, the proposed system ensures that the age and gender demographics will be reflected proportionally in the sequence of advertisements that are about to be presented.

The rest of this paper is organized as follows: The next section introduces the notion of a group TV profile in the context of age and gender demographics, and how existing audio analysis techniques could be combined to construct a group profile. Section III introduces a genetic-based recommender, extended to computing of 7-dimensional age and gender ratings, and section IV demonstrates how an M-user group profile can be adapted to an N-slot advertisement profile, and show how this is used to drive a genetic algorithm-based recommendation engine. Following this the experimental setting is discussed. The system is then evaluated from a number of perspectives. Finally, conclusions are drawn.

II. Group Profile Derivation

Solving the ”Who is sitting in front of the TV?” problem is a challenging task and has yet to be researched fully. When only one person watches TV, attempting to derive additional profile attributes by means of speech or an acoustical analysis does not make much sense, and instead one must rely on other sources of information, such as an explicitly provided user profile, or through image recognition (many households today already have movement detection cameras as a standard games console accessory).

A typical system could be realized as follows: When multiple users are present, the audio from several microphone pickups in the room is applied to an independent component analysis algorithm that can separate the background TV audio (if any) from the users' speech [9]. Speaker diarization is used on the speech part to separate speaker utterances of different people from one another, and to determine the number of speakers present at any given time [10], [11].

In the ideal case, where the exact age and gender class for each user in the audience is known, a Group Viewer Configuration (GVC) is formed, which can be expressed as follows:

$$GVC = \begin{bmatrix} c_{user_1} \\ c_{user_2} \\ \vdots \\ c_{user_M} \end{bmatrix} \tag{1}$$

where c_{user_m} corresponds to the age and gender class of the *m-th* user, $1 \le c \le C$, C is the total number of age and gender classes, $1 \le m \le M$ and M is the total number of users.

In practice, the speaker utterances from each speaker in the audience are classified according to age and gender to determine their class. However, due to the probabilistic nature in which speaker classification systems work, along with their limited accuracy, it is important to note that each speaker, regardless of their age/gender class, will to some extent be a member of *all* other classes. In this way, the user profile for a single user m whose real class is c_{user_m}, can be modeled by:

$$\boldsymbol{x_m} = \begin{bmatrix} p_{m,1} \\ p_{m,2} \\ \vdots \\ p_{m,C} \end{bmatrix} \tag{2}$$

where $p_{m,j}$, $0 \le p_{m,j} \le 1$, simply represents the actual predicted probability for class j, $1 \le j \le C$, and $\sum_{j=1}^{C} p_{m,j} = 1$. The more utterances that can be collected, the better the classification accuracy.

For all M users, a group profile is then constructed from the individual user profiles as follows:

$$\boldsymbol{X_G} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_M^T \end{bmatrix} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,C} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \cdots & p_{M,C} \end{pmatrix} \tag{3}$$

III. Demographic Recommendation

The matching problem can be stated as optimizing the match between the group profile and the sequence of content items (advertisements) that are about to be presented to the users. The basic genetic algorithm approach proposed for extending MacauAp [12] is taken as the starting point for the recommender system, and performs the relevant matching. Based on user-feedback of categories for tourist destinations, called ”spots”, the genetic algorithm in MacauAp searches amongst a large number of tourist destinations and finds a sequence with an optimum match between the categories to which the spots belong and the user-liked categories. In the same vein, the purpose of the genetic algorithm for the proposed system is to find the sequence of content items, whose combined demographic profile best matches the audio-derived group profile.

Genetic Algorithms are established computational methods that conduct their searches based on natural selection and genetics, and use the concepts of chromosomes, populations, selection, crossover and mutation [13]. A search is typically initiated by creating a population comprised of units called chromosomes, where each chromosome is effectively a sample of the search space. Each iteration of the algorithm entails selecting two parents from the population (selection) using a tournament or proportionate selection approach. A fitness evaluation is conducted to determine which of the chromosomes ought to be considered as parents. From these parent chromosomes a child chromosome (crossover) is then spawned that comprises attributes from both parents. The child then replaces the weakest chromosome in the population. This process continues until termination, which is usually defined as the point where the population becomes stable. The chromosome with the highest fitness value is then selected as the winner, and is the output of the algorithm. In this approach a chromosome is defined as a sequence of content items (advertisements) that are to be broadcast in the next upcoming break. Each chromosome has N slots, where a slot is defined as a placeholder for a single content item.

Before parents can be selected for crossover, the fitness of each chromosome needs to be computed. As was proposed for the MacauAp scenario, the base fitness for such a chromosome is given as

$$Fitness_{base} = \sum_{i=1}^{N} r_i * Pref_i \tag{4}$$

where

$N =$ Number of slots in the chromosome,
$r_i =$ Official rating for slot i, and
$Pref_i =$ User preference for slot i.

The scalar rating r_i from equation (4) is extended in this work by converting it to a vectorized form where all age and gender classes are represented:

$$\boldsymbol{r_i} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_C \end{bmatrix} \tag{5}$$

The predefined age and gender ratings for the N content items (equal to the number of slots) is then given as:

$$\begin{bmatrix} r_1^T \\ r_2^T \\ \vdots \\ r_N^T \end{bmatrix} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,C} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N,1} & r_{N,2} & \cdots & r_{N,C} \end{pmatrix} \tag{6}$$

For now each user is assumed to be assigned to a single slot, making $M = N$, meaning that each user in the GVC gets to see at least one content item to his/her liking (shortly this will be extended to the case $M \neq N$). Treating $Pref_i = \boldsymbol{x_m}$, where user m is assigned to slot i, now allows one to express the fitness more compactly as:

$$Fitness = \mathrm{tr}(R * X_G^T) \tag{7}$$

where $\mathrm{tr}(A)$ is the trace of A, i.e. the sum of the main diagonal of A.

IV. Group Profile Adaptation

Equation (7) above requires that for each slot, there is a separate set of preferences values. Since however it cannot be assumed that $M = N$ (for example, there might be 4 users present, but 5 advertisements are to be presented), a new set of preferences with the same dimension as N is needed. The intention is to ensure that each user's demographic membership for all age and gender classes is carried over proportionally to the new preference set. This adaptation process is best explained using the double circle diagram shown in Figure 1, which shows the adaptation process for a single age/gender class[2]. The circle is divided into a fixed number of bins, which run at equally spaced intervals over its entire revolution. On the outer circle one allocates an equal portion of the bins to each user, so for example, for 4 users, there will be 4 separate partitions. The same applies to the inner circle, but instead of allocating bins to users, they are allocated according to slots. For 5 slots one therefore ends up with 5 equally-sized slot partitions. It therefore follows intuitively that for the bins comprising a single slot, rating contributions can come from multiple users. The amount that each user contributes to a given slot is directly proportional to the size of the overlap between the user bins and slot bins. Summing over all bins belonging to a single slot partition and dividing by the number of bins per slot, allows one to compute a rating for that slot.

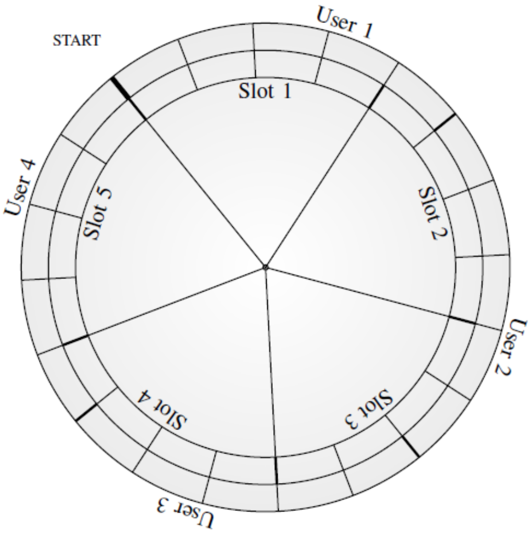

Fig. 1. Proportional bin selection for a single age and gender class, for 4 users, 5 slots and 20 bins. The rating for a given slot can come from multiple users.

More formally, assuming the number of bins is B, a new rating matrix $\bar{\Sigma}_{C,N}$ is defined, where each element $\bar{\Sigma}_{i,j}$ is given as:

$$\bar{\Sigma}_{i,j} = \frac{1}{\mu}\,\bar{X}_{G_{i,\phi(j,k)}} \tag{8}$$

where

$$B = LCM(M, N),\ \text{i.e. the least common multiple,} \tag{9}$$
$$\mu = \frac{B}{N},\ \text{i.e. slot partition size in bins, and} \tag{10}$$
$$\nu = \frac{B}{M},\ \text{i.e. user partition size in bins} \tag{11}$$

and where

$$\phi(j,k) = \left\lfloor \left(\frac{(j-1)*\mu+k-1}{\nu}\right) + 1 \right\rfloor \tag{12}$$

represents the user partition to which bin k that is currently being processed, belongs to.

Now armed with separate ratings for each slot, the fitness can now finally be calculated as:

$$Fitness = \sum_{i=1}^{N}\sum_{j=1}^{C}\bar{R}_{i,j} * \bar{\Sigma}_{i,j} \tag{13}$$

where

$N =$ Number of content items (slots) to present,
$C =$ Total number of age and gender classes,
$\bar{R}_i =$ Normalized rating of class j for slot i, and
$\bar{\Sigma}_{i,j} =$ Normalized group membership for class j for slot i.

V. Experimental Setup

A. Group Viewer Configurations

To emulate the home group viewers, 50 separate GVCs are defined, which are shown in Table I. The viewer configurations that were chosen were based on information provided by Statistics Denmark, which records comprehensive statistics on the composition of Danish households. Here it can be seen that 23.8 % of the population live alone, 38.7 % live with one other person, 14.3 % belong to a family of three, 14.6 % belong to a family of four and 5.5 % belong to a family of five. From these figures, and based on the fact that single-person profiles are excluded, the viewer configurations selected are based on families of two, three and four persons, where the bulk of the distribution lies.

Now just for the two-person households, children and youngsters do not feature much, and only comprise 2.8 % and 2.1 % of households, respectively. In contrast, 37.6 % of households contain adults and 57.5 % have seniors, giving the first ten configurations in Table I below.

Looking at children and youth from just the three- and four-person households, it can be noted that for children, 30.1 % are part of three-person families, but that 69.4 % (more than double) are part of four-person families. For the youth category, 40.1 % of youths belong to three-person families whereas 59.9 % of youths belong to four-person families. Thus it is evident that children and youths should feature fairly strongly in the chosen configurations. The data also shows that there were twice as many seniors with two children living at home (15657 people) than seniors with only one child living at home (7302). Finally, a number of four-person configurations spanning all generations are included. From all this, the two-person configurations 11-17, the three-person configurations 18-30 and the four-person configurations 31-50 in Table I below[3] are constructed.

[2] The adaptation for each age and gender class is computed independently of the others.

[3] Note that one differentiates between the number of people in the household, and the number of viewers in front of the TV, e.g. there will be multiple two-person and three-person configurations for a four-person household.

TABLE I
SELECTED TV VIEWER CONFIGURATIONS

No.	Profile	Profile	Profile
1-3	AM|AM	AM|AF	AF|AF
4-6	SM|SM	SM|SF	SF|SF
7-9	AM|SM	AM|SF	AF|SM
10-12	AF|SF	C|C	C|YM
13-15	C|YM	C|AM	C|AF
16-18	C|SM	C|SF	C|C|SM
19-21	C|C|SF	C|SM|SF	C|C|YM
22-24	C|C|YF	C|C|AM	C|C|AF
25-27	C|YM|YF	C|YF|YF	C|YM|YF
28-30	C|AM|AM	C|AM|AF	C|AF|AF
31-33	C|C|C|YM	C|C|C|YF	C|C|C|AM
34-36	C|C|C|AF	C|C|C|SM	C|C|C|SF
37-39	C|C|C|C	C|C|YM|YM	C|C|YM|YF
40-42	C|C|YM|AM	C|C|YF|AF	C|C|AM|AF
43-45	YM|YM|AM|AF	C|YM|AM|AF	C|AM|AF|SM
46-48	C|AM|AF|SF	C|C|SM|SF	YM|YM|SM|SF
49-50	YM|YF|SM|SF	AM|AF|SM|SF	

Selected TV group viewer configurations. C=Child, YM=Young Male, YF=Young Female, AM=Adult Male, AF=Adult Female, SM=Senior Male, SF=Senior Female

B. Audio Classification of Age and Gender

For each speaker from the viewer configuration profile, a set of real speaker utterances are classified to determine their class. The utterances are selected by randomly picking out a speaker with a matching class from the evaluation portion of the aGender corpus [14]. The aGender corpus was supplied to participants in the InterSpeech 2010 Paralinguistic Challenge to enhance the development of age and gender algorithms. The training part of the dataset contains 32527 utterances from 472 speakers, the development part contains 20549 utterances from 300 speakers and the testing part contains 17332 utterances. It comprises 4 age classes: children (7-14 years), young people (15-24 years), adults (25-54 years) and seniors (>= 55 years), and 3 gender classes: children, males and females. Children are classed as their own gender since males are indistinguishable from females at that age. In more recent work, the age boundaries are slightly different, i.e. children (<= 13 years), young people (14-19 years), adults (20-54 years) and seniors (>= 55 years) [6]. The latter age boundaries, corresponding to the recent work, were chosen[4].

For the speaker that was selected, the speaker utterances for the speaker were pooled to form a contiguous segment. Each speech segment was then submitted for classification, to determine its class. The speaker results were then combined to form the group profile $\boldsymbol{X_G}$ from above.

The audio classification system is constructed as a hybrid system comprising two subsystems. The first subsystem models acoustic speaker features and the second subsystem models the prosodic features. Modeling several feature types increases the classification accuracy of the system.

The acoustic subsystem is modeled using the well-known Gaussian Mixture Model Universal Background Model (GMM-UBM) approach [15]. After voice activity detection was applied to each utterance [16], feature extraction was performed using 13-dimensional Mel Frequency Cepstral Coefficients (including C0), with 1st and 2nd derivative, to give 39 coefficients per 25 ms frame (15 ms overlap). Mel Frequency Cepstral Coefficients (MFCCs) are simply a compact representation of the spectral envelope of a speech signal. A 512-component GMM-UBM was trained using all the training data from the aGender corpus. Seven speaker models, one for each class, were then adapted from the single UBM using the training data from each class. For the adaptation process, a relevance ratio of 12 was used. The accuracy for the acoustic subsystem for all classes was 49.9 %.

To model the prosody features the prosody baseline referred to as System 7 in a previous work was used [6], which models prosody features at the syllable level instead of at the frame level. The syllable boundaries are determined as follows: For each utterance, all frames are marked as voiced or unvoiced (unvoiced where the pitch is undefined) and all unvoiced frames are discarded. For the remaining frames, the normalized energy contour is used as a key to determining the syllable boundaries, where valleys in the contour indicate the start of a new syllable.

The prosody features modeled for each syllable are contours of pitch, energy, formants, syllable duration and spectral harmonic energy (obtained from the power spectrum at harmonics of F0). The Praat package was utilized to extract pitch and energy features from each utterance and Matlab was used to compute the spectral harmonic energy. After applying time scale normalization for the interval -1 to 1, the contours were then modeled as sixth-order Legendre polynomials, meaning that instead of an entire contour, only six coefficients need to be stored [17]. Seven 512-component GMM models were then trained with the prosody features, one for each class. The prosodic subsystem's accuracy for all classes was 42.0 %.

For the hybrid system, the acoustic and prosodic subsystems were combined using weighted summation-based fusion [6] of the subsystem results. The hybrid classifier model was tested on the entire development data set, an accuracy was achieved on the combined system of 50.0 %. As a comparison, another work using seven individual subsystems was able to attain an accuracy of 50.3 % [6][5].

C. Initial Rating of Advertisements

The advertisement corpus used in this paper was provided courtesy of TV2, a nationwide television broadcaster in Denmark. The commercials are subdivided into 24 categories. Examples of categories are *Food*, *Beverages*, *House and Home* and *Media*.

[4] The original aGender age boundaries were chosen solely on the basis of marketing aspects, and not on any physiological aspects.

[5] Whereas the hybrid system appears to only offer a marginal increase in accuracy over the acoustic-only system, it should be borne in mind that this represents the average for all 7 age and gender classes for each system, and that the response for individual classes for the two systems are in some cases quite different.

A random subset of advertisements was taken from each category to give a total of 200 advertisements. These were then split into four separate groups. For each group of 50 advertisements, three subjects were asked to rate 50 of them, on the basis of how well they matched the seven age and gender categories listed above. The scale used was the standard 1-5 likert scale, with 1 being not-relevant and 5 most relevant. The ratings were then averaged for each advertisement across the participants for each group, by taking the median.

Table II shows a sample selection of the rated advertisements.

TABLE II
SELECTED COMMERCIALS

Ad	C	YM	YF	AM	AF	SM	SF
Washing machine	1	3	3	5	2	4	1
Computer game	4	5	3	4	1	2	1
Alcohol drink	1	3	5	1	4	3	2
Lift chair for the elderly	1	1	1	1	1	4	5
Children's building bricks	4	4	2	3	2	1	1
Ferry company	2	2	1	5	5	5	4
Retail bank	1	2	3	5	4	1	1

Commercial details of advertisements have been withheld.

D. Recommendation of Advertisements

The matching of advertisements was carried out as follows: The advertisements were combined from the four rating groups to give a total of 200 advertisements. The genetic algorithm was initialized with 50 chromosomes, with each chromosome comprising N randomly chosen advertisements. It was not possible for the same advertisement to appear twice within a given chromosome. After this, 500 iterations of genetic selection were run, where the fitness in each round was recomputed according to the predefined advertisement's age and gender ratings and the extracted group profile. At the end of the entire run, the chromosome with the highest fitness was selected as the winner and used as the sequence of recommended advertisements.

VI. Evaluation and Results

A. Evaluation of User Categories

In the first part of the evaluation, a system using an audio-derived group profile (the proposed system) is compared to an ideal system, where the group profile matching a given GVC is provided explicitly, for example through an online user form or registration questionnaire.

In the proposed system an audio classifier determines the M-user group profile by connecting real speaker utterances to each user in the GVC and computing a probabilistic membership for each class. In the ideal system, there is no audio classification, and instead each user profile $\boldsymbol{x_m}$ is constructed by setting a value of 1 for the class matching the user category, and a 0 for the remaining classes. The users are combined in the same way to form the group profile.

The objective for either system is for a given GVC, to recommend a sequence of content item where for each user, a single matching advertisement is suggested that has the same age and gender category. Since in this case each user maps to a single ad, $M = N$ and no group profile adaptation is carried out in this particular evaluation. The recommended category for an advertisement is simply defined as the advertisement's age and gender class that received the highest rating. Some of the advertisements, however, had two or more classes that had received the same, highest rating. In these cases the advertisements were assigned to multiple categories.

The overall effectiveness of the system in predicting the correct category was measured as follows: for each advertisement that was recommended, an input-output pair was formed containing the true user category, e.g. *Child*, and the category of the advertisement that was recommended, e.g. *Young female*. To take account of the advertisements that had multiple categories, additional input-output pairs were created as necessary. For each recommendation made, there is therefore at least one mapping from input category to output category. For each GVC, 50 evaluations were carried out, with different speaker utterances being used in each evaluation (with possible replacement).

Once all 50 GVCs had been tested, the input-output pairs were collated and entered into a 7x7 contingency table. To measure the level of association between the input categories (the rows, R) and the output categories (the columns, C), Pearson's chi-squared test was carried out. The strength of association using Cramer's Phi (or V) for categorical variables can be computed for each system as:

$$\phi_c = \sqrt{\frac{X^2}{T(k-1)}} \tag{14}$$

where

$X^2 =$ Pearson's chi-square statistic,
$T =$ Total number of input-output pairs and
$k = \min(R, C)$.

Here T is computed as:

$$T = Sim_{count} * GVC_{count} * N_{gvc} * Cat_{Count} \tag{15}$$

where

$Sim_{count} =$ Number of group simulations,
$GVC_{count} =$ Number of GVCs,
$N_{gvc} =$ Number of content items(ads) recommended, and
$Cat_{count} =$ Number of highest rated categories for each item.

Furthermore the value of k is 7, since $R = C = 7$. The obtained values for each system are shown in the Table III and Table IV below.

TABLE III
CONTINGENCY TABLE FOR IDEAL SYSTEM

Cat	C	YM	YF	AM	AF	SM	SF	Sub
C	3175	747	626	19	25	1	2	4945
YM	288	761	553	231	152	32	34	2051
YF	61	236	446	96	255	19	57	1170
AM	1	69	59	899	480	395	185	2088
AF	0	39	135	399	850	146	315	1884
SM	0	20	20	519	239	698	388	1884
SF	0	16	55	280	550	406	749	2056
Sub	3525	1888	1894	2443	2551	1697	1730	15728

Explicit group profile. $X^2 = 18293, \phi_c = 0.4403, T = 15728$. Key: *Cat* is Category, *Sub* is row or column subtotal

TABLE IV
CONTINGENCY TABLE FOR PROPOSED AUDIO SYSTEM

Cat	C	YM	YF	AM	AF	SM	SF	Sub
C	2585	1452	1698	367	559	115	183	4595
YM	98	537	440	587	442	284	246	2634
YF	186	234	381	118	247	45	96	1307
AM	28	344	291	841	558	585	374	3021
AF	87	210	422	396	755	276	495	2641
SM	16	171	157	661	439	614	400	2458
SF	60	152	261	443	667	407	583	2573
Sub	3060	3100	3650	3413	3667	2326	2377	21593

Audio-extracted group profile. $X^2 = 9295, \phi_c = 0.2678, T = 21593$. Key: *Cat* is Category, *Sub* is row or column subtotal

The results show that the effect size $\phi_c = 0.2678$ of the audio-based system is lower than that of the theoretical maximum given by the ideal system of $\phi_c = 0.4403$, which can be accounted for by the error introduced by the audio classifier. Depending on what speaker utterance was classified, there will always be varying degrees to which a given class is a member of the other classes (never 0 % or 100 %). This can also clearly be seen in the diagonals of each table, which represent the number of hits correctly classified for each user category. In the ideal system, there are (as expected) a larger number of correctly classified advertisements.

What is interesting to note for the proposed system, however, is that there is a significant correlation between the input categories and output categories. This is in spite of the accuracy of the audio classifier only being half that of the theoretical maximum. To test for significance, the null hypothesis may be stated that the age and gender categories of the advertisements for each user are chosen with equal probability. With a goodness of fit of $X^2 = 9295$ and $\text{degrees of freedom} = 36$, it was found that with a $p < 0.1$, that the null hypothesis is disproved. Furthermore, the value of $\phi_c = 0.2678$ for the proposed system is over 0.25, which according to the threshold values for Cramer's V, corresponds to a very strong association between the true categories and the recommended categories.

What is also interesting to note in the proposed system is the value of T, which happens to be 27 % higher than in the ideal system, meaning more multi-category advertisements were selected in the proposed system. The reason for this is believed to be the classifier-induced increased membership of each class of every other class, which leads to selecting advertisements with multiple highly-rated classes.

B. Evaluation of Group Adaptation

When group-to-slot adaptation is performed, $M \neq N$ and the number of advertisements to be recommended is different to the numbers of users sitting in front of the TV. Since the number of advertisements for a given break are predetermined, a system that does not employ the adaptation technique would have to resort to other methods to fill up the additional slots. In this part of the evaluation two systems are compared: a system where the remaining $N - M$ slots are filled with random advertisements and a system where the full group profile adaptation technique is applied.

To formally evaluate to what extent each user's age and gender class is represented in the sequence of recommended items, and to determine whether group profile adaptation gives a better proportional representation than the alternative system, an adapted version of the group profile X_G is correlated with the age and gender ratings of the recommended advertisements. The idea behind this is that the more accurately the individual user's classes are reflected in the sequence of items that is presented, the stronger the correlation will be. For example, if the initial GVC was [C|C|C|SF], meaning that three quarters of the audience are children, then it is expected that three quarters of the advertisements will also be targeted to children.

To allow for this comparison, each GVC is converted to a 7-dimensional age-and gender representation $\boldsymbol{x_{gvc}}$, where a 1 is given for each age-and-gender class in the GVC, and a 0 is given otherwise. In doing so, the class that has the strongest representation in the GVC will end up having the largest weighting. Likewise a similar 7-dimensional age and gender representation $\boldsymbol{x_{items}}$ is constructed for the N advertisements that are represented. To determine the weighting for each class, each advertisement's ratings for all classes are summed across each individual class. In this way, the class that has the strongest representation across all N items will end up having the largest weighting.

To compute the correlation between $\boldsymbol{x_{gvc}}$ and $\boldsymbol{x_{items}}$, Kendal's Rank Correlation Coefficient is used[6], which is a non-parametric hypothesis test that measures the degree of concordance between the values being compared. The test ensures that before the strength of the correlation is computed, the data on both sides is ranked, and where tie ranks are observed, the rank value simply becomes the average of the individual ranks. For each τ_B value that is computed, the corresponding z-score is also computed, which is characterized by a normal distribution when the variables are statistically independent.

[6] Kendal's τ_B is considered a better statistic for smaller amounts of data than Spearman's Rank Coefficient, and where the ranks that are to be compared have ties.

The results show that there is a stronger overall effect, and hence preservation of the original age and gender classes, when group profile adaptation is applied ($\tau_B = 0.2316$, $Z_B = 34.29$) than when random advertisements are used to fill the remaining slots ($\tau_B = 0.08$, $Z_B = 11.91$). The values for τ_B and Z_B were also calculated for each of the 50 GVCs for both schemes. When the direction of correlation is taken into account (zero correlation is considered better than a negative correlation), it was noted that in 47 out of 50 cases that a stronger rank correlation coefficient (and accompanying z-score) was obtained for the case where group profile adaptation was employed. Therefore, from a statistical standpoint, there is a stronger overall effect introduced when applying adaptation. Table V shows the values for the first 10 individual τ_B and Z_B values.

TABLE V
EFFECTIVENESS OF APPLYING GROUP ADAPTATION

P	$\boldsymbol{\tau_B(1)}$	$\boldsymbol{\tau_B(2)}$	$\boldsymbol{Z_B(1)}$	$\boldsymbol{Z_B(2)}$	P	$\boldsymbol{\tau_B(1)}$	$\boldsymbol{\tau_B(2)}$	$\boldsymbol{Z_B(1)}$	$\boldsymbol{Z_B(2)}$
1	0.44	0.47	8.99	9.44	2	0.52	0.55	10.49	11.12
3	0.45	0.44	9.06	8.94	4	0.11	0.38	2.16	7.76
5	0.13	0.23	2.68	4.71	6	0.12	0.31	2.34	6.24
7	0.41	0.55	8.21	11.20	8	0.25	0.32	5.10	6.42
9	0.29	0.31	5.88	6.34	10	0.11	0.27	2.25	5.51

Key: P=Profile, $\tau_B(1)$ and $Z_B(1)$ corresponds to System 1 (random ads in remaining slots) and $\tau_B(2)$ and $Z_B(2)$ corresponds to System 2 (full adaptation). Shown for the first ten GVCs.

C. *User Study Evaluation*

In another evaluation relating to this study [18], a user study was conducted where subjects were given the chance to evaluate the proposed system. Due to time constraints, it was not possible to evaluate all 50 group configurations, and therefore a subset of 11 configurations was selected for the study. A total of 12 subjects were asked to participate in the study.

Subjects were shown several different GVCs, and for each GVC, a sequence of 10 advertisements. Five of the advertisements were recommended using a system where the hybrid audio classifier, group profile adaptation algorithm and genetic selection algorithm were present in the system. The other 5 advertisements were randomly selected, without replacement. Subjects were however, not informed which advertisements were recommended and which were randomly selected. They were then asked to rate on a scale of 1-10 (1 completely irrelevant and 10 most relevant) on whether they thought a given advertisement was suitable for any of the members of the GVC. A 10-point scale was used to ensure that subjects took a non-neutral stance when rating. For example if the subject thought the advertisement appealed highly to children, and the child category was part of the GVC, then naturally the advertisement would receive a higher rating.

The results show that on average the recommended advertisements received a higher median rating of 7.75 than the randomly selected ones, which received a rating of 4.25. To test the statistical significance of the recommended items receiving higher ratings, let x represent all samples corresponding to the median ratings of all users for the randomly-selected items and let y represent samples corresponding to the median ratings of all users for recommended items, and test the null hypotheses that $y - x$ comes from a distribution of zero median. Treating the rating scales as ordinal, the 2-sided Wilcoxen Signed Rank test is used to test for significance. With a z-score $z = -2.628$ and $p < 0.01$ there is a significant increase in the median rating for the recommended items, thus disproving the null hypothesis. Finally, the effect size using Pearson's correlation coefficient $r = \frac{Z}{\sqrt{N}}$ is also computed, where Z is the z-score from above and $N = 24$ is the number of observations, and is found to be $r = 0.535$. Since the absolute value is above Cohen's benchmark of 0.5, it can be concluded that using the age-and-gender analysis approach has a large effect on the user ratings.

VII. CONCLUSION AND FUTURE WORK

This paper showed how an audio classifier can be used to elicit a demographic group profile from a given audience, and how this can be used to provide recommendations. Even at the level of state-of-the-art age and gender detection, which is about 50 %, there is good potential in using audio analysis for recommendation. For the proposed system it was found that there was a strong relationship between the true user categories and the recommended advertisement categories. In the majority of cases, group profile adaptation leads to a stronger reflection of the users' age and gender classes than simply adding random advertisements to the remaining slots. User studies confirm that the strength of the recommendation can be perceived and that the recommended advertisements were more suitable than randomly selected advertisements.

Finally, it is proposed that the system be used as a baseline for future work. This includes an investigation into further novel ways in which the accuracy of detecting the age and gender of viewers can be enhanced, with the intention to see to what extent it is possible to approach the upper bound results for the explicitly-provided group profile system.

VIII. ACKNOWLEDGMENT

The authors wish to thank TV2 Denmark for providing the video TV commercials and Felix Burkhardt for supplying the aGender corpus.

REFERENCES

[1] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee, "AIMED - a personalized TV recommendation system," *Lecture Notes in Computer Science*, vol. 4471, pp. 166–174, 2007.

[2] Y.-C. Chen, H.-C. Huang, and Y.-M. Huang, "Community-based program recommendation for the next generation electronic program guide," *IEEE Transactions on Consumer Electronics*, vol. 55, pp. 707–712, 2009.

[3] M. Z. Bjelica, "Unobtrusive relevance feedback for personalized TV program guides," *IEEE Transactions on Consumer Electronics*, vol. 57, pp. 658–663, 2011.

[4] D. Bonnefoy, M. Bouzid, N. Lhuillier, and K. Mercer, "”More Like This” or ”Not for Me”: Delivering personalised recommendations in multi-user environments," *Lecture Notes in Computer Science*, vol. 4511, pp. 87–96, 2007.

[5] C. Shin and W. Woo, "Socially aware TV program recommender for multiple viewers," *IEEE Transactions on Consumer Electronics*, vol. 55, pp. 927–932, 2009.

[6] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, pp. 151–167, 2012.

[7] T. I. Meinedo H., “Age and gender detection in the I-DASH project,” *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, 2011.

[8] A. Maier, J. G. Bauer, F. Burkhardt, and E. Nth, “Age and gender recognition for telephone applications based on GMM supervectors and support vector machines,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1605–1608, 2008.

[9] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurocomputing*, vol. 41, pp. 1–24, 2001.

[10] Z.-H. Tan, “Audio and speech processing for data mining,” *Encyclopedia of Data Warehousing and Mining - 2nd Edition*, vol. 1, pp. 98-103, 2008

[11] S. E. Tranter, “An overview of automatic speaker diarization systems,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1557–1565, 2006.

[12] R. Biuk-Aghai, S. Fong, and S. Yain-Whar, “Design of a recommender system for mobile tourism multimedia selection,” *Internet Multimedia Services Architecture and Applications (IMSAA)*, 2008.

[13] G. D., “Genetic and evolutionary algorithms come of age,” *Communications of the ACM*, vol. 37, pp. 113–119, 1997.

[14] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, “A database of age and gender annotated telephone speech,” *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*, pp. 1562–1565, 2010.

[15] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[16] Z.-H. Tan and B. Lindberg, “Low-complexity variable frame rate analysis for speech recognition and voice activity detection,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 798–807, 2010.

[17] N. Dehak, P. Dumouchel, and P. Kenny, “Modeling prosodic features with joint factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2095–2103, 2007.

[18] S. Shepstone, Z.-H. Tan, and S. H. Jensen, “Demographic recommendation by means of group profile elicitation using speaker age and gender detection,” *Accepted for publication at Interspeech 2013, Lyon France*, 2013.

BIOGRAPHIES

Sven Ewan Shepstone received the B.S. and M.S. degrees in Electrical Engineering from the University of Cape Town in 1999 and 2002 respectively. He is employed at Bang and Olufsen A/S in Denmark and is currently an industrial PhD candidate at Aalborg University. His main research interests are digital TV and the application of speech technologies to recommender systems.

Zheng-Hua Tan 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. He is an Associate Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark, which he joined in May 2001. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include speech and speaker recognition, noise robust speech processing, multimedia signal and information processing, multimodal human-computer interaction, and machine learning. He has published extensively in these 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/serves as a program co-chair, area and session chair, tutorial speaker and committee member in many major international conferences.

Professor Søren Holdt Jensen 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. Before joining the Department of Electronic Systems of Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd, Copenhagen, Denmark; the Electronics Institute of the Technical University of Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education, Lyngby; the Electrical Engineering Department of Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK) of Aalborg University. He is Full Professor and is currently heading a research team working in the area 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 an 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 member of the Danish Council for Independent Research - Technology and Production Sciences by the Danish Minister for Science, Technology and Innovation.