Demographic Recommendation by means of Group Profile Elicitation Using Speaker Age and Gender Recognition

Sven Ewan Shepstone[1], *Zheng-Hua Tan*[2], *Søren Holdt Jensen*[2]

[1]Bang and Olufsen A/S,
Peter Bangs Vej 15, 7600 Struer, Denmark
[2]Department of Electronic Systems, Aalborg University,
Niels Jernes Vej 12, 9220 Aalborg, Denmark
ssh@bang-olufsen.dk, zt@es.aau.dk, shj@es.aau.dk

Abstract

In this paper we show a new method of using automatic age and gender recognition to recommend a sequence of multimedia items to a home TV audience comprising multiple viewers. Instead of relying on explicitly provided demographic data for each user, we define an audio-based demographic group profile that captures the age and gender for all members of the audience. A 7-class age and gender classifier employing a fusion of acoustic and prosodic features determines the probability of each speaker belonging to each class. The information for all speakers is then combined to form the group profile, which itself is the input to a recommender system. The recommender system finds the content items whose demographics best match the group profile. We tested the effectiveness of the system for several typical home audience configurations. In a survey, users were given a configuration and asked to rate a set of advertisements on how well each advertisement matched the configuration. Unbeknown to the subjects, half of the adverts were recommended using the derived audio demographics and the other half were randomly chosen. The recommended adverts received a significantly higher median rating of 7.75, as opposed to 4.25 for the randomly selected adverts.

Index Terms: age identification, gender identification, demographic filtering, acoustic and prosodic fusion, genetic algorithms, group recommendation

1. Introduction

This paper shows how a state-of-the-art age and gender classifier can be leveraged to power a recommender system for selecting TV content. Instead of basing the age and gender profile needed for recommendation on manually provided data or usage patterns, we propose using audio analysis methods instead.

The detection of age and gender is a complicated task and has received a lot of research interest recently. Typically, the age and gender of speakers are identified by means of Gaussian mixture models, multilayer perceptrons, hidden Markov models and/or support vector machines [1], [2]. In particular, modern age and gender classification results are making it more and more feasible to use on-the-fly demographic classification for recommendation purposes. The state-of-the-art accuracy of gender-only classifiers is roughly 30 % higher than that of age detection [3]. The same work shows that a system using automatic speaker recognition using a fusion of acoustic and prosodic features was able to achieve an accuracy of 85.0 % for the gender classification task, 52.0 % for the age classification task and an accuracy of 50.3 % for the combined age and gender classification task [3].

What is interesting to note is that the largest confusion occurs between speakers of the same sex (e.g. young males, adult males and senior males) and between children and young females. While there is still room for future improvement, we believe that there is a strong basis for recommendation, since the effect of overlapping confusion classes could well be ameliorated by soft preference and market boundaries. For example, with respect to short advertisement clips, there are many products that would appeal to both young males and adult males, or to both children and young females, thus canceling out some of the effects of the confusion overlap between these classes.

Collaborative recommender systems are the most widespread recommender systems in use today and rely on a large user base of ratings to make recommendations. Essentially, these systems work by correlating the feedback rating of a user for a specific item with that of other users for the same item, to make recommendations for a new item that is unknown to the user (but that the others rated) [4]. However, with home set-top boxes there is no easy way to exchange the user ratings, with the result that for these types of systems, a content-based approach is more applicable [5].

Content-based recommenders can determine similarities directly between content items and a given user profile, provided the user profile can be extracted, and there exists suitable metadata for content items[1]. However, the need for a user profile implies that the profile must either be explicitly provided, for example by means of a questionnaire when registering a set top box [6], or implicitly, by building the profile by monitoring usage patterns [5]. A bigger problem, however, is when multiple consumers share a single device, such as a home television, but each has their own user profile and tastes [7]. This occurs often with home game playing and movie watching, where typically only one username or profile is utilized.

Our contribution in this paper is a novel method of using audio analysis techniques to extract the parameters needed for constructing a group profile for recommendation. This is in contrast to traditional methods of using user questionnaires, usage data or ratings to collect the viewers' data. We focus primarily on age and gender in this study, and utilize an age and gender

This work is supported by the Danish Ministry of Science, Innovation and Higher Education under Grant no 12-122802.

[1]Collaborative systems and content systems are often deployed in a hybrid configuration to take advantage of their strengths.

classifier to provide a group demographic profile for communal TV viewing. We test the hypothesis that given a particular home viewer configuration, and given a group profile derived using an audio analysis of each member of the configuration (audience), that the recommended items (advertisements) will receive higher ratings from users, than if the content items were randomly selected, thus indicating a closer match to the viewer configuration[2].

The remainder of the paper is as follows: Section 2 introduces the notion of a demographics-based audio group profile. We then discuss adapting the group profile to make it usable for recommendation. Section 4 presents the home viewer configuration used in this study, and the audio classifier that transforms a viewer configuration to a group profile. We then discuss experimental work and the surveys that were conducted. Finally we present our results and draw conclusions.

2. Extracting the Audio Group Profile

Solving the "Who is sitting in front of the TV?" problem is challenging and has yet to be researched fully. A typical system could be realized as follows: The audio from several microphone pickups in a room could be applied to an independent component analysis algorithm that separates the background TV audio (if any) from the users' speech [8]. 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 [9], [10]. The speaker utterances from each speaker can then be classified according to age and gender, which in turn can be used to construct a group profile. Due to the limited accuracy of current state-of-the-art age and gender systems, it is important to note that each speaker, regardless of their age and gender class, will to some extent be a member of all defined age and gender classes. In this study, motivated by the corpus that was used for training our classifier [11] and by recent works [3], [1], we base our study on seven such classes.

The user profile for each speaker m, generated over a set of utterances for that speaker, can be modeled by:

$$x_m = \begin{bmatrix} p_{m,1} \\ p_{m,2} \\ \vdots \\ p_{m,C} \end{bmatrix} \quad (1)$$

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

For a set of M users, we then define a group profile as:

$$X_G = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,C} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \cdots & p_{M,C} \end{pmatrix} \quad (2)$$

3. Matching and Recommendation

The matching problem can be stated as optimizing the match between the group profile X_G, obtained by classifying a set of utterances for each speaker, and the sequence of content items (ads) that the viewers will see. When the number of users M is equal to the number of items N we allocate *one* item per viewer, thus allowing each viewer to see a content item of their liking.

When $M \neq N$ (N might be fixed, due to e.g. the length of an ad break) there is no longer a 1-1 mapping between users and items. In this case we perform what we refer to as group profile adaptation. This entails converting the group profile X_G, which represents M users, to a new profile Y_G, which represents N pseudo-users, and where N is now equal to the number of items to present. This means that for each class in the original group profile, we determine the proportional membership of each user to that class. For example, assume a 2-user group profile that must be extended to 3-pseudo users. For the first class (Child), we find that the first user has a 80 % membership of the class (leaving only 20 % to all other classes), while the second user has a 40 % membership of the class (leaving 60 % to all other classes). For 3 pseudo-users, we split the pseudo-user space up into 3 equally-sized portions. The first pseudo-user overlaps completely with the 1st user - hence it receives an 80 % membership. The 2nd pseudo-user overlaps $50 - 33.3 = 16.7$ % with the 1st user and $66.6 - 50 = 16.7$ % with the 2nd user. The membership for this pseudo-user is then proportionally calculated as $\frac{80*16.7+40*16.7}{16.7+16.7}\% = 60$ %. Finally since the third pseudo-user overlaps completely with the 2nd user, we just assign the same membership of user 2 to the third pseudo-user, i.e. 40 %. Note that when $M = N$ then $Y_G = X_G$.

Now for a given content item

$$c_n = \begin{bmatrix} p_{n,1} \\ p_{n,2} \\ \vdots \\ p_{n,C} \end{bmatrix} \quad (3)$$

which has a predefined age and gender profile, the strength of the match for each user-item pair is then simply computed as:

$$Match_{n,n} = Y_G(n, *) * c_n \quad (4)$$

To perform the actual matching we use a modified form of genetic algorithm, proposed previously for providing itinerary-based recommendations [12]. 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].

Upon initialization, the algorithm selects k chromosomes, each containing N randomly-chosen ads. The strength of each chromosome (how well it matches the adapted group profile) is then computed by taking the sum of content-item matches, with each match computed as shown in Equation 4 above. With each iteration of the algorithm, the chromosome with the poorest match to the adapted group profile is discarded, and replaced with a new genetically-spawned sequence.

For our experiments, the ad selection process was as follows: We first initialized our genetic algorithm with $k = 50$ chromosomes of 5 ads each. The ads were taken from a central pool of 200 ads and it was not possible for an ad to appear twice within a given chromosome. We then ran 500 iterations of genetic selection, and selected the sequence with the strongest match likelihood as the sequence of ads to be recommended.

4. Age and Gender Audio Classification

4.1. Viewer Configuration Profile

To emulate a group containing several viewers of varying demographics, we define a viewer configuration profile. To select

[2]We do not evaluate the system using prediction error, since there is no ground truth (all ads rated for every group viewer configuration).

which viewer configurations to use, we turned to Statistics Denmark [14], which records comprehensive statistics on the composition of Danish households. Here we could see 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, we based our viewer configurations 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 don't 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 configurations 1 and 2 in Table 1 below.

Looking at children and youth from just the three- and four-person households, we note 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 our chosen configurations. From this, we construct configurations 3, 4, 5, 6, 7, 8 and 9 shown in Table 1 below[3].

Finally we examined statistics on the number of seniors ($\geq$ 55) with children and/or youngsters living at home. We found 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), giving the last two configurations.

Profile No	1	2	3	4
Profile	AM+AF	SM+SF	C+C	C+YM
Profile No	5	6	7	8
Profile	C+YF	C+AM	C+AF	C+C+AM
Profile No	9	10	11	
Profile	C+C+AF	C+SM	C+SF	

Table 1: TV viewer configuration. C=Child, YM=Young Male, YF=Young Female, AM=Adult Male, AF=Adult Female, SM=Senior Male, SF=Senior Female

This gives a total of 11 configurations. For each configurations that was presented (explained below), we broke it up into its constituent parts, i.e. individual speakers, and for each speaker, connected them to real speaker utterances.

4.2. Dataset

The speaker utterances used for classification were taken from the aGender corpus, which was supplied to participants in the InterSpeech 2010 Paralinguistic Challenge to enhance the development of age and gender algorithms [11]. 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 ($\geq$ 55 years), and 3 gender classes: children[4], males and females. In more recent work, the age boundaries are slightly different, i.e. children ($\leq$ 13 years), young people (14-19 years), adults (20-54 years) and seniors ($\geq$ 55 years) [3]. We chose to use the latter age boundaries from the recent work.[5]

[3]In this study we focus on 2 and 3 people at a time in front of the TV.

[4]Children are classed as their own gender since males are indistinguishable from females at that age.

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

4.3. Speaker Classification

For each speaker from the viewer configuration profile we randomly selected a speaker with the matching class in the evaluation portion of the aGender dataset. To represent this speaker we pooled together the selected speaker's utterances 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 $\bar{X_G}$ from above.

For classification we employed a hybrid system, where each age and gender class is modeled separately. Both acoustic and prosodic features are modeled, with fusion of acoustic and prosodic features occurring at the utterance level.

The GMM baseline was constructed using the well-known UBM-GMM approach [15]. After voice activity detection [16], feature extraction was performed using 13-dimensional MFCCs (including C0, 1st and 2nd derivative), to give 39 coefficients per 25 ms frame (15 ms overlap). We then trained a 512-component GMM UBM using all the training data from the aGender corpus. Following this, 7 speaker models were adapted from the UBM using the training data from each class. For the adaptation process, we used a relevance ratio of 12. The accuracy for the acoustic sub-system for all classes was 49.9 %.

To model the prosody features we used the prosody baseline referred to as System 7 in a previous work [3], and which models prosody features at the syllable level instead of 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). We used the Praat package [17] to extract pitch and energy features from each utterance and Matlab 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 [18]. We then trained 7 512-component GMM models with the prosody features, one for each class. The accuracy for the prosodic subsystem for all classes was 42.0 %.

We then combined the two acoustic and prosodic subsystems together in a hybrid system using weighted summation-based fusion [3] of the sub-system results. We tested our hybrid classifier model on the entire development data set, where we achieved an accuracy on the combined system of 50.0 %. As a comparison, another work using seven individual sub-systems was able to attain an accuracy of 50.3 % [3]. A more detailed breakdown of the 2 classifiers is shown in Table 2 below.

5. Experimental work

The advertisement corpus used in this paper has 24 categories of ads and was provided to us courtesy of TV2, a Danish public-service television broadcaster. To be able to match advertisements with the group profile discussed above, we conducted a pre-survey to annotate each ad with an age and gender profile. We took a random subset of ads from each category, giving a total of 200 commercials, which we then split into four separate groups. For each group of 50 ads, three subjects were asked to rate all 50 commercials, on the basis of how well they

	C	YM	YF	AM	AF	SM	SF
C	**69.6**	3.4	16.2	1.7	4.4	1.3	3.5
	61.0	7.5	16.9	2.0	4.9	1.0	6.7
YM	1.6	44.8	1.3	27.4	0.3	19.7	4.9
	0.3	**49.4**	0.8	21.9	1.0	23.5	3.2
YF	18.7	2.2	49.9	1.3	21.6	0.5	5.7
	16.4	0.8	**57.1**	0.3	15.8	0.6	9.0
AM	2.3	20.7	0.3	**47.8**	1.3	25.1	2.5
	0.1	29.2	0.0	27.1	1.1	40.5	2.2
AF	10.4	3.5	21.1	1.9	**40.2**	1.0	21.9
	5.5	1.8	26.6	0.4	33.8	0.6	31.3
SM	2.5	14.5	0.2	23.6	0.5	55.9	2.8
	0.2	11.5	0.1	16.2	0.2	**69.7**	2.0
SF	10.5	4.7	11.6	2.1	24.9	4.3	41.9
	7.1	1.5	11.4	0.9	22.9	2.2	**53.9**

Table 2: Confusion matrix for seven-class Age and Gender Classifier. Shaded entries are the results for our classifier (two sub-systems; overall accuracy 50.01). Non-shaded entries are the results of a recent work (seven sub-systems; overall accuracy 50.3). **Bold** typeface shows the better score of the two systems.

thought each ad matched all seven age and gender classes. The scale used was the standard 1-5 Likert scale (1 not-relevant and 5 most relevant). For each ad rated by three separate people, we took the median rating for each class as the official rating for the advertisement. Table 3 shows a sample selection of ads, with their corresponding median ratings.

Short Description of Ad	C	YM	YF	AM	AF	SM	SF
Women's sandals	1	1	5	2	5	1	4
Cleaning Agent	1	1	4	3	5	1	5
Lift Chair	1	1	1	1	1	5	5
Chewing Gum	1	5	4	4	4	4	4
Dating Site	1	1	1	5	5	2	2
Hair Product	1	3	5	1	5	1	5
Chocolate Easter Egg	5	1	1	3	3	1	2
Building Blocks	5	1	1	2	5	2	3

Table 3: Selected ads with accompanying ratings.

To test the effectiveness of using the acquired audio group profile in our recommender, we conducted another survey where subjects were shown a set of home viewer configurations, and for each configuration, asked to rate a set of 10 advertisements. A different set of advertisements was used for each round. For each set shown, five of the ads were obtained by using the genetic algorithm approach and the other five were randomly selected (without replacement) from an initial pool of 200. The set was then shuffled before being presented for recommendation. Subjects were not told that five of the ads for the given slot had been randomly selected, thus giving them no way of knowing which of the ads had been recommended. They were then asked to rate each ad on a scale of 1-10 (1 completely irrelevant and 10 most relevant), on the basis of the ad appealing to *any* of the members of the home viewer configuration. For example, if the subject thought the ad appealed highly to children, and the *Child* category was part of the configuration, then the ad would receive a higher rating. A 10-point scale was used to ensure that subjects took a non-neutral stance when rating.

We used 12 subjects for our evaluation. Since it was not possible time-wise for each subject to rate all 11 proposed viewer configurations, we split the configurations into 3 groups. The first four subjects were therefore asked to evaluate the first four group viewer configurations, the second four subjects were asked to rate advertisements for the second four configurations, and the last four subjects were asked to rate advertisements for the last three configurations.

6. Results

We now look at the results that were obtained. Table 4 shows two median ratings for each user of the survey. The first rating was taken as the median of all ratings performed for the user on the randomly selected ads, whereas the second rating was taken as the median of all ratings for the recommended ads.

Test Subject	Random	Recommended
1	7	9.5
2	2.5	7
3	10	8
4	7	9
5	4.5	10
6	3	5
7	4	5
8	4.5	7.5
9	4	8
10	4	10
11	2	7
12	8	7

Table 4: Average ratings for the 12 users, taken for the random case and recommended case. Average for each user taken using the median.

From the averages in Table 4 we see that the recommended ads obtained consistently higher ratings than the random ads. Only 2 of the users (users 3 and 12) returned an average rating for the random ads that was higher than the recommended ads.

To test the statistical significance of the recommended ads receiving higher ratings, we let x represent all samples corresponding to the median ratings of all users for the random ads and y represent samples corresponding to the median ratings of all users for recommended ads, and test the null hypotheses that $y-x$ comes from a distribution of zero median. Treating the rating scales as ordinal, we use the 2-sided Wilcoxen Signed Rank test to test for significance. We find with a z-score $z = -2.628$ and $p < .01$ that there is a significant increase in the median rating for each test group, thus disproving the null hypothesis. Indeed, from the table above, the user ratings for the recommended group have a median of 7.75, which was significantly higher than the ratings for the random group, with a median of 4.25. We also compute the effect size using Pearson's correlation coefficient $r = \frac{Z}{\sqrt{N}}$, where Z is the z-score from above and $N = 24$ is the number of observations, and find it to be $r = -0.535$. Since the absolute value is above Cohen's benchmark of 0.5, we can conclude that using the age-and-gender analysis approach has a large effect on the user ratings.

7. Conclusion

This paper showed how an age and gender classifier using mixed acoustic and prosodic features can be used to elicit a demographic group profile from a given audience, and how this can be used to provide recommendations. The classifier we built delivered comparable results to the state-of-the-art and showed that there is a basis for recommendation, even with large gaps of confusion between classes. We showed that ratings for adverts recommended using the age and gender data were significantly higher than ratings for randomly selected adverts.

8. Acknowledgments

The author wishes to thank Bang and Olufsen A/S for sponsoring this research, TV2 Denmark for providing the video TV commercials and Felix Burkhardt for supplying the aGender corpus.

9. References

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

[2] 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 Acoustics, Speech and Signal Processing*, pp. 1605–1608, 2008.

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

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, pp. 734–749, 2005.

[5] A. B. Barragns-Martneza, E. Costa-Montenegroa, J. C. Burguilloa, M. Rey-Lpezb, F. A. Mikic-Fontea, and A. Peleteiroa, "A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition," *Information Sciences*, vol. 180, pp. 4290–4311, 2010.

[6] 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.

[7] 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.

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

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

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

[11] 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.

[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] "Statistics denmark." [Online]. Available: http://www.statistikbanken.dk/statbank5a/default.asp?w=1680

[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] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.4.32) [computer program]," 2009, available from http://www.praat.org/.

[18] 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.