Age-Group and Gender Classification through Class-Dependent Phone Recognition

Michael Norris, Michael Wagner

National Centre for Biometric Studies, University of Canberra, Australia
michael.wagner@canberra.edu.au, michael.norris@canberra.edu.au

Abstract

This study proposes a method to determine the gender and age group of a speaker by means of an automatic speech recognition system that is trained on six different sets of phones: one for each intersection of the two gender and three age-group classes. The study uses the Australian National Database of Spoken Language (ANDOSL) with 18 speakers in each class reading a set of 200 phonetically rich sentences. The system trains 44 context-independent phone models for each of the six classes and determines the gender and age group of an unknown utterance by finding the best matching phone sequence against the combined set of 264 phone models. Two methods of utilising the resulting phone sequences for gender and age-group recognition are evaluated: firstly, simple counting of the number of phones that belong to each class is used as the basis for the six-way class decision; secondly, the recognised phone sequence is converted to a 264-dimensional vector, whose components contain the phone counts in the phone sequence for each of the 6 x 44 phones in the combined set. An artificial neural network is trained to make the final gender and age-group decision using the count vectors as input. The artificial neural network outperforms the simple counting method with an average correct recall for gender of 97.7%, an average correct recall for age group of 60.5% and an average correct recall for combined gender and age group of 58.9%.

Index Terms: Gender classification, age classification, speaker characterisation.

1. Introduction

A five-year collaborative project of four Australian universities and four partner research groups in Germany, Denmark and the USA is currently underway in Australia, named "From Talking Heads to Thinking Heads" [1]. The purpose of the project is to enhance the intelligence of the human-machine interface of a system, which presents as a talking avatar with audio and video sensing capability, and to enable the system to engage human interlocutors in interesting dialogue.

The current study is a part of the effort to add speaker characterisation capability to the Thinking Head system and to connect the system's ability to detect characteristics such as gender, age, dialect, accent and emotional state of the human interlocutor with a capacity of the system's dialogue management subsystem to adapt its dialogue strategy to the specific characteristics of the human it is presently speaking with. For example, the system could change a number of aspects of its behaviour, from choice of words, syntax and conversation topic to its facial gestures, depending on whether it is responding to a 20-year old woman or a 60-year old man, or to a happy customer compared with an angry one.

This study addresses the detection of age and gender in particular, a problem that has been studied by a number of different researchers in recent years (see [2] for a recent overview). In our case the motivation springs from the general observation that human listeners are very good at determining a speaker's gender and reasonably good at estimating, at least roughly, the age or age group of a speaker [2]. In order perhaps to pass the Turing Test [3] or at least to provide reasonably intelligent conversation with a human, we are therefore endeavouring to build such speaker characterisation capability into the Thinking Head system.

A secondary motivation is to achieve speaker characterisation within the existing structures of our system instead of building a separate subsystem for that purpose. We are therefore showing here that it is possible to achieve good estimates of gender and age by utilising the automatic speech recognition capability that is already part of the Thinking Head system.

An opportunity for this study is provided by the Australian National Database of Spoken Language (ANDOSL). ANDOSL contains a set of 200 phonetically rich sentences of Australian English, read by 108 different speakers who are divided into two gender groups and three age groups [4]. Each combined gender-age group comprises 18 speakers, which has led to our experimental design of training a separate phone set for each of the six classes and then detecting, to which of the six classes an unknown utterance belongs.

The mechanism we use was inspired by other work in the areas of speaker, language and accent identification by means of phone recognition [e.g. 5]. We use a set of 44 standard phones of Australian English, but we train acoustic models for those phones separately for each of the six groups of speakers, yielding a combined set of $6 \times 44 = 264$ acoustic phone models such as 44 elderly male models, 44 middle-aged female models etc. We then test an utterance by letting a phone recogniser find the best-matching phone string, given the combined set of 264 phones. The resulting phone string, which in general contains a mix of male phones and female phones, and elderly, middle-aged and young phones, is analysed and leads to a score against each of the six speaker classes. The scores are derived in different ways: either by counting the proportions of phones in the phone string that belong to the six different classes, or by processing the counts for all the 264 phones present in that string through an artificial neural network (ANN) for numerical scores against the different classes. The partitioning of the data corpus was done in such a way that two thirds of the speakers were used to build the acoustic models, one sixth to develop the weights of the ANN and one sixth for the testing and determination of average recall rates [6] for the gender and age classes.

In the following Sections, we will first describe our speech data corpus in Section 2.1. Next, we will, in Section 2.2., detail the speech recognition system, including the tagged-phone inventory, the dictionary, and the acoustic models. In

ISBN 978-0-9581946-3-1

Section 2.3., we describe the decoding process and the resulting strings of tagged phones, and in Section 2.4. the ANN is detailed, which transforms the tagged-phone sequence into decision scores for the gender and age-group detection. Section 3 presents the results of the study, followed by some conclusions in Section 4.

2. Experimental Design

2.1. Corpus

The Australian National Database of Spoken Language (ANDOSL)[1] is currently the only sizeable publicly available speech data corpus of Australian English. It contains read speech and spontaneous speech by native speakers as well as by accented speakers. This project uses the read speech by each of 108 native speakers of Australian English of 200 phonetically rich sentences. The 108 speakers are grouped into 18 groups of 6 speakers. The 18 groups divide into 2 genders (female and male), three age groups (elderly, middle-aged and young), and three sociolects (cultivated, general and broad). The young group consists of speakers aged between 17 and 30 years, the middle-aged group consists of speakers aged between 31 and 45 years, and the elderly group consists of speakers aged between 46 and 70 years. The sociolect was determined at the time of the ANDOSL recordings by an expert phonetician. The ANDOSL audio data were recorded at a sample rate of 20 kHz and at 16 bits/sample. In order to be compatible with the Carnegie Mellon University Sphinx 3 [7] automatic speech recognition system, the audio was downsampled to 16 kHz for this study.

For this study we combined the three sociolect groups and thus considered the database as comprising 6 groups – 2 genders × 3 age groups – with each group having 18 speakers. The 18 speakers in each group were partitioned into 12 speakers for training and 3 speakers each for development and testing. All partitions are balanced with respect to sociolect. The acoustic models for the automatic speech recognition were then trained with the 12 training speakers for each group. Then, the weights of the ANN were trained with the 18 speakers of the development set (6 groups × 3 development speakers), and finally the entire age and gender detection was tested with the remaining 18 speakers of the test set (6 groups × 3 test speakers). The development and test phases were duplicated by swapping the development speakers and the test speakers.

2.2. Acoustic model training and tagged phones

During the training phase, the CMU Sphinx 3 automatic speech recognition system [7] was used to obtain a sequence of tagged phones for each training sentence. Acoustic models were trained for the phone set in each class using the forced alignment feature of Sphinx. Phone transcriptions were derived from the known orthographic transcriptions of the corpus by means of a pronunciation dictionary, such that where a word has several possible pronunciations in the dictionary, the one that best fits the audio is chosen. Lacking a specific Australian-English pronunciation dictionary, a combination of the Cambridge British English Example Pronunciations (BEEP) [8], the Australian subset of the Edinburgh Unisyn dictionary [9] and several smaller pronunciation dictionaries were used, mapping the orthography of the sentences to sequences of phones from a 44-phone inventory.

Phone transcriptions for each speaker were tagged with the gender and age group of that speaker, producing transcriptions from a combined inventory of 264 phones: 44 speech sounds each tagged with a gender: M/F and age group: Y/M/E. For example, the phone set includes the phone FEEA (female, elderly, phone /EA/) and MYT (male, young, phone /T/), and each phone in the transcription for a female, middle-aged speaker was tagged with the prefix "FM". Tagged phones provide a mechanism for information about the speaker to be obtained conveniently from an ASR engine such as Sphinx.

The 264 acoustic models were trained using the CMU Sphinx tool SphinxTrain. The models are context-independent (CI) monophones because the number of possible triphones would have been prohibitively large. Each CI model has three states. The dictionary for this task was simply a one-to-one mapping of each tagged phone to its output representation.

2.3. Decoding

Decoding of the test and development sets was performed using the CMU Sphinx 3 decoder as a context-free phone recogniser.

The tagged phone sequences resulting from phone recognition do not represent accurate word or sentence recognition. However, the characteristics of the speaker in terms of both gender and age group were expected to be correlated usefully with the recogniser's choice of phones tagged with the speaker's gender and age group over other phones.

Table 1. *Means and standard deviations for the proportions of recognised phones belonging to the correct speaker characteristic (gender or age group)*

Class	Proportion of correct gender tags		Proportion of correct age tags	
	mean	std	mean	std
FE	0.856	0.120	0.365	0.237
FM	0.867	0.104	0.430	0.114
FY	0.900	0.072	0.417	0.127
ME	0.852	0.116	0.494	0.129
MM	0.790	0.138	0.289	0.113
MY	0.790	0.113	0.363	0.099
Avg	0.843	0.111	0.393	0.137

The statistics of the proportion of occurrence of correctly tagged phones in decoded sentences are summarised in Table 1 for gender and age group, separately. Each row of the table is a speaker group. Columns of the table are the means and standard deviations of the proportions of in-group tags recognised in sentences. The first pair of columns are the means and standard deviations of the proportions of tags of the correct gender. For example, the number in the top-left corner says that 85.6% of the phones in all female-elderly test data were correctly tagged as female, i.e. as FE, FM or FY. The average proportion of correct gender tags is 84.3%, which compares with the chance level of 50%. The second pair of columns are the means and standard deviations of the proportions of tags of the correct age group. For example, the third number in the top row says that 36.5% of the phones in all female-elderly test data were correctly tagged as elderly,

[1] Note, however, that the collection of a new Australian speech corpus is undertaken collaboratively by a consortium of Australian Universities in 2010.

i.e. as FE or ME. The average proportion of correct age-group tags is 39.3%, which compares with the chance level of 33.3%.

Table 2 shows similarly the means and standard deviations of the proportion of tags of the correct gender-AND-age group recognised. For example, the number in the top-left corner says that 30.4% of the phones in all female-elderly test data were correctly tagged as female-elderly, i.e. as FE only. The average proportion of correct gender-and-age-group tags is 33.3% compared with a chance level of 16.7%.

Table 2. *Means and standard deviations for the proportions of recognised phones belonging to the correct speaker characteristic (gender and age group)*

Class	Proportion of correct gender-age tags	
	mean	std
FE	0.304	0.199
FM	0.376	0.125
FY	0.381	0.139
ME	0.403	0.126
MM	0.229	0.097
MY	0.304	0.107
Avg	0.333	0.132

It can be seen from the tables that there is a preponderance of tagged phones from the correct gender-age group. These results then motivated us to investigate a suitable method to convert the proportions and identities of the recognised phonemes into scores for the recognition of the gender and age group.

2.4. Decision networks

To optimise the decision on the correct gender-and-age group, rather than just using a simple threshold, we used the statistics of occurrence of the 264 phones in each sentence to decide the gender and age-group for the sentence. A 264-bin histogram was built for each decoded sentence, recording the frequency of occurrence of each tagged phone. These 264-element vectors were then used as input to a multilayer perceptron [10]. Two networks were trained, each having 264 inputs, 2 outputs and 3 hidden units. The two outputs represent gender (female- not female) and age-group (young-not young in one network, elderly-not elderly in the other). Tagged phones from the decoded sentences in the development data set were used to train the neural network.

The roles of the development and test data partitions were then reversed to provide a second example of the experiment.

3. Results

The development data were processed through the ANN in several iterations until convergence and the connection weights determined by way of the back propagation algorithm. The detection threshold was then optimised for the equal-error-rate criterion – obtaining equal error rates for false acceptance and for false rejection –, and finally the performance of the system was determined on the basis of the 18 speakers × the 200 sentences of the test data partition.

The tables show the number of correct recalls for each class: in the first column for the correct recall of gender, in the second column for the correct recall of age group, and in the last column for the gender-age-group combined class. Table 3 shows a two-group classification between the young age group and the two older age groups, while Table 4 shows a two-group classification between the elderly age group and the two younger age groups

The results in the Tables are averaged over two runs where the development set and the test set of data are exchanged. The gender recall for the two different ANNs varies between 95.0% and 100.0% with a mean of 97.7%. The age-group recall varies between 41.6% and 79.2% with a mean of 60.5%. In both cases, those results compare with a chance level of 50%. The combined gender-AND-age-group recall varies between 41.3% and 77.1% with a mean of 58.9%, compared with a chance level of 25%. It is distinctly noticeable that the recall rates for the middle-aged male and female groups are lower than the corresponding elderly and young age groups. This is likely to be a result of the way the two ANNs each makes a binary decision between elderly and not-elderly on the one hand and between young and not young on the other.

Table 3. *Correct class recall rates for the two genders and a young-vs-not-young class boundary*

Class	Gender recall	Age-grp recall	Combi recall
FE	0.950	0.613	0.591
FM	0.988	0.416	0.413
FY	0.999	0.574	0.573
ME	0.990	0.735	0.726
MM	0.989	0.511	0.504
MY	0.966	0.792	0.771
Avg	0.980	0.607	0.596

Table 4. *Correct class recall rates for the two genders and an elderly-vs-not-elderly class boundary*

Class	Gender recall	Age-grp recall	Combi recall
FE	0.951	0.531	0.486
FM	0.990	0.463	0.460
FY	1.000	0.603	0.603
ME	0.993	0.722	0.718
MM	0.978	0.511	0.496
MY	0.928	0.788	0.723
Avg	0.973	0.603	0.581

4. Conclusions

We have presented a gender-and-age classification for Australian English by way of a flat phone recogniser and a phone inventory of 264 class-dependent phones, followed by a multilayer perceptron to optimise the age and gender decision. The method has the advantages of obtaining the recognised phone sequences for gender and age detection from the existing ASR system of our Thinking Head system. Further analysis is planned to obtain additional information from the artificial neural networks, which will point to differences between the phones in their capacity to distinguish between the sexes and the different age groups.

5. References

[1] "From Talking Head to Thinking Head"-website: http://thinkinghead.edu.au/, retrieved on 27.02.2010.

[2] Schötz, S., Acoustic analysis of adult speaker age, in C. Müller [ed] Speaker Classification I, pp 88-107, Springer 2007.
[3] Turing, A.M., Computing machinery and intelligence, Mind, Vol. LIX, No. 236, October 1950.
[4] ANDOSL website. http://andosl.anu.edu.au/andosl/, accessed 23.07.2010.
[5] Schultz, T., Jin, Q., Laskowski, K., Tribble, A. and Waibel, A., Speaker, Accent, and Language Identification Using Multilingual Phone Strings, Proceedings of HIT 2002, Second International Conference on Human Language Technology Research, M. Marcus, ed., Morgan Kaufmann, San Francisco, 2002.
[6] Makhoul, J., Kubala, F., Schwartz, R. and Weischedel, R., Performance measures for information extraction. in Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999.
[7] Carnegie Mellon Sphinx-website: http://cmusphinx. sourceforge.net/
[8] Pronunciation dictionary, http://svr-www.eng.cam.ac.uk/~ajr/wsjcam0/node8.html, accessed on 23.07.2010.
[9] Unisyn Lexicon, http://www.cstr.ed.ac.uk/projects/unisyn/, accessed on 23.07.2010.
[10] Haykin, S, Neural Networks: A Comprehensive Foundation (2nd ed.). Prentice Hall, 1998.