Information Bottleneck Based Age Verification

Ron M Hecht [1], *Omer Hezroni* [1], *Amit Manna* [1], *Gil Dobry* [2], *Yaniv Zigel* [2], *Naftali Tishby* [3]

[1] PuddingMedia, Kfar-Saba, Israel
[2] Bio-medical Engineering Department, Ben-Gurion University, Beer-Sheva, Israel
[3] School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel
hadasron@gmail.com, omer.hezroni@puddingmedia.com, amit.manna@puddingmedia.com, gil.dobry@gmail.com, yaniv@bgu.ac.il, tishby@cs.huji.ac.il

Abstract

Word N-gram models can be used for word-based age-group verification. In this paper the Agglomerative Information Bottleneck (AIB) approach is used to tackle one of the most fundamental drawbacks of word N-gram models: its abundant amount of irrelevant information. It is demonstrated that irrelevant information can be omitted by joining words to form word-clusters; this provides a mechanism to transform any sequence of words to a sequence of word-cluster labels. Consequently, word N-gram models are converted to word-cluster N-gram models which are more compact. Age verification experiments were conducted on the Fisher corpora. Their goal was to verify the age-group of the speaker of an unknown speech segment. In these experiments an N-gram model was compressed to a fifth of its original size without reducing the verification performance. In addition, a verification accuracy improvement is demonstrated by disposing irrelevant information.

Index Terms: age verification, age estimation, speech processing, information bottleneck.

1. Introduction

The N-gram word model is a commonly used model in different speech recognition tasks. It is used as a language model for large vocabulary continuous speech recognition (LVCSR) engines in full transcription tasks. In addition, it is used in lexical-based speaker verification [1] and age estimation [2] for secondary scoring of transcriptions. A similar N-gram model is used for all of these tasks regardless of the task's specific goal. Although being an effective model, it has a significant drawback. By the time that an N-gram model contains ample amount of information X, only a small fraction of that is relevant to the classification goal Y. A model with substantial amount of irrelevant information has several disadvantages; the most significant being the relatively large amount of training data needed to train it compared to a model that contains only relevant information. In the last few years several approaches have mitigated this disadvantage. Such approaches include reducing the number of words in the model, maintaining a low N-gram level (mainly unigrams and bigrams) or even by using different features such as Term Frequency–Inverse Document Frequency (TF-IDF) [3][4][5][6]. In this paper, a novel approach to the extraction of *relevant* information is presented. It enables the ability to distinguish among different age groups using the Information Bottleneck method [7].
The Information Bottleneck (IB) method attempts to find a compact model representation $\hat{X}$ to replace X, such that it preserves as much of the relevant information Y as possible. This method was already used to tackle other challenges such as relevant speech feature extraction [8], document categorization [9] and analysis of neural codes [10].
In this paper, the IB method is applied for a word-based age-group verification task. An age-group verification that is based on the transcription of an LVCSR engine is proposed. This system is constructed of three consecutive stages. In the first stage a N-gram is estimated and is followed by two classification stages: a support vector machine (SVM) stage followed by a log likelihood ratio (LLR) stage. The N-gram estimation stage is divided into two steps; first a word N-gram model is estimated and then a compact representation of the model is achieved by using the Agglomerative Information Bottleneck method (AIB) [11]. It is shown that by applying the AIB method an improved model was achieved: one that is more compact and preserves a significant amount of the relevant information. Consequently, the age verification performance using this method was improved.

2. Methods

In this section, the Information Bottleneck method and one of its greedy derivatives, the Agglomerative Information Bottleneck method, are described.

2.1. Information Bottleneck Method

The IB method enables us to find a more compact model representation $\hat{X}$ by providing a tradeoff between two contradicting objectives. On one hand, the IB preserves as much of the relevant information as possible by maximizing the mutual information I between the compact model representation $\hat{X}$ and the classification goal Y.

$$I\left(Y;\hat{X}\right)=\sum_{\substack{\hat{x}\in\hat{X}\\ y\in Y}} p(y,\hat{x})\log\frac{p(y,\hat{x})}{p(y)p(\hat{x})} \tag{1}$$

where $p(y,\hat{x})$ is the joint probability of the goal and the compact representation. In this paper for example, it might be the joint probability of the word unigram clusters and the speakers' ages.
On the other hand, the IB discards the irrelevant information by minimizing the mutual information between the original model representation X and the compact model representation $\hat{X}$. The tradeoff between the two objectives is known as the Information Bottleneck functional.

$$L\left[p\left(\hat{x}|x\right)\right]=I\left(X;\hat{X}\right)-\beta I\left(Y;\hat{X}\right) \tag{2}$$

Where β is a positive Lagrange multiplier.

2.2. Agglomerative Information Bottleneck Method

The AIB is a greedy approximation to the minimization of the IB functional. This derivative of the IB is applicable for a discrete feature space and is used for bottom-up hierarchical clustering. The algorithm starts with a set of clusters such that for every object in the discrete feature space a cluster is created and therefore generates a number of clusters equivalent to the number of objects. In the unigram model for example, the algorithm starts with a set of clusters such that every cluster has a single word in it. In each step the algorithm starts with a given clustering $\hat{X}_{before}$ and merges two objects x_l, x_r to a new object x_m according to the AIB criterion, resulting in a new clustering $\hat{X}_{after}$. The criterion for the selection of the clusters is to minimize the loss of L:

$$x_m = x_r \cup x_l = \\ = \arg\min_{x_r' \cup x_l'} \left(L\left[p\left(\hat{x}_{before}|x\right)\right] - L\left[p\left(\hat{x}_{after}|x\right)\right] \right) \quad (3)$$

Since only x_r, x_l, x_m contribute to the change in L, the former equation can easily be solved and yields the following merge criterion.

$$\begin{aligned} & L\left[p\left(\hat{x}_{before}|x\right)\right] - L\left[p\left(\hat{x}_{after}|x\right)\right] = \\ & = I\left(x;\hat{x}_{before}\right) - \beta I\left(y;\hat{x}_{before}\right) - \left(I\left(x;\hat{x}_{after}\right) - \beta I\left(y;\hat{x}_{after}\right)\right) = \\ & = p(x_r) KL\left(p\left(x|x_r\right)\|p(x)\right) + p(x_l) KL\left(p\left(x|x_l\right)\|p(x)\right) \\ & - p(x_m) KL\left(p\left(x|x_m\right)\|p(x)\right) - \beta p(x_r) KL\left(p\left(y|x_r\right)\|p(y)\right) \\ & - \beta p(x_l) KL\left(p\left(y|x_l\right)\|p(y)\right) + \beta p(x_m) KL\left(p\left(y|x_m\right)\|p(y)\right) \end{aligned} \quad (4)$$

where KL is the Kullback–Leibler divergence.

$$KL\left(p(x)\|q(x)\right) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (5)$$

3. AIB-Based System

Based on the AIB algorithm a word-based age-group verification system was created. This system takes advantage of the ability of the AIB algorithm to extract relevant information. The system has three training phases and a single testing phase.

Training phase *A* is the most fundamental phase where word-level clustering occurs. The input for this stage is a large corpus in which each of the age groups is sufficiently represented. The output is a mapping between the word-level features, word unigrams for example, and clusters. The phase is composed of several stages. In the first two stages the word based features are created. During the first stage, the audio is transformed to a sequence of words using an LVCSR engine. In the second stage, the word sequence is transformed to word level features. The word level features used are word unigrams $p(w_i)$ and word pairs unigrams $p(w_i, w_{i-1})$. In each of the experiments only one type of feature is used. During the third stage the joined distribution between the age groups and the word-level features is estimated. The word-level AIB-based clustering occurs during the fourth stage. The AIB algorithm clusters the features into groups until it reaches a desired number of clusters.

The goal of phase *B* is to train an SVM model for the classification of the different age groups. Each conversation in the corpus is first mapped to a point in the clusters distribution space, where each dimension of the space is the conditional probability of the i^{th} cluster given the j^{th} call $p\left(x_i|c_j\right)$ where c_j denote the j call. This is achieved by first transforming the audio to word-level feature as in phase A. The mapping from training phase A is then used to transform the word-level feature to cluster distributions. The input conversations of this phase are the input for the SVM training. For each age-group a single one-versus-all SVM model is trained.

During the last training phase *C*, a LLR model is trained. For each age-group a Gaussian model with a diagonal covariance matrix is estimated. During this stage each call is transformed to a point in the clusters distribution space in a manner identical to the first three stages of training phase B. Each point is then scored against all the SVM models that were trained in the previous stage. The output is a point in the SVM score space where the dimension of the space is the number of models used. In this space the LLR model is estimated.

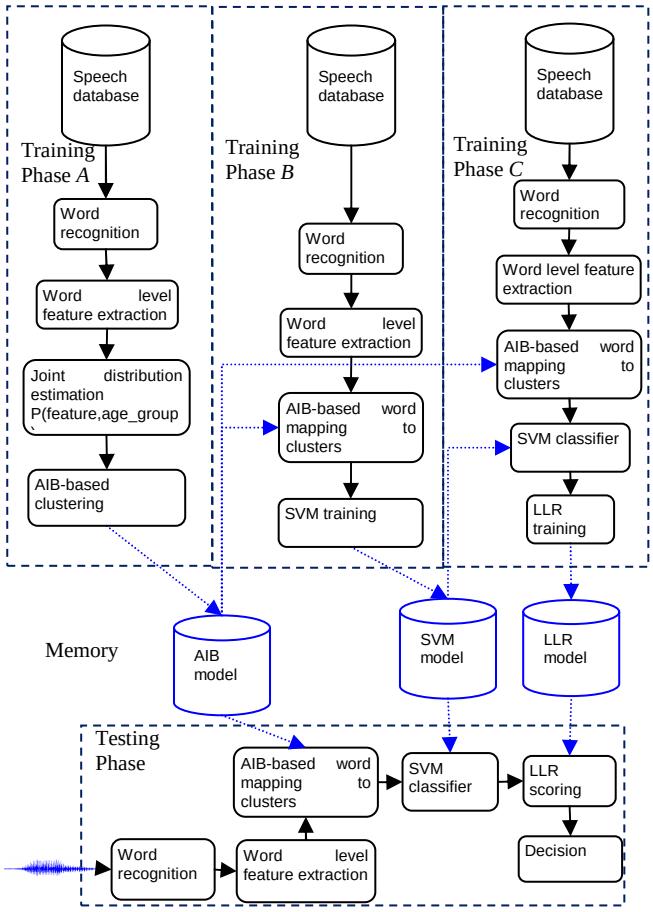

Figure 1: *The AIB-based age estimation system.*

The testing phase takes advantage of all the model types that were estimated in the training phases. The testing phase includes the first four stages of training phase C. At the last stage, the LLR stage, the Gaussian with the diagonal covariance matrix for the hypothesized age group is used for the final scoring.

4. Experiments and Results

This section describes two sets of experiments that were conducted on the LDC Fisher English speech corpora [12]. The goal of the first set of experiments was to find the most natural split of the age spectrum into age-groups. The second set of experiments was a set of age-group verification experiments. These experiments demonstrated the advantages of the AIB method. The task that was chosen was a four age-groups verification task where the age group of a speaker for a given conversation is verified. Each side of a conversation was treated as a different call.

4.1. Corpora

The Fisher corpora were selected due to their vast number of calls and speaker diversity. Thousands of calls were available for each age group, each around ten minutes long, providing tens of thousands of minutes of relevant audio. These calls were recorded from thousands of distinct speakers. In addition, the corpora are well balanced between both genders. Approximately 11,000 conversations from the Fisher part II corpus were used for training while tests were conducted on more than 8,000 conversations from the Fisher part I. Conversations in the original Fisher part I corpus were omitted from the tests if the speaker of a conversation was present in the Fisher part II corpus.

Table 1. Training and Testing Corpora Sizes

Age-group	Training	Testing
0-25	2595	1894
26-40	4202	3309
41-55	2853	2155
56+	1114	868

4.2. Age Group Definition

In most speech recognition challenges, the challenge is defined as the classification among well defined and distinct groups. In gender recognition, for example, the caller is either male or female. However in age estimation, it is much more complicated to draw the borders between age groups due to several reasons:

- The target variable is a continuous one.
- A person does not spend all his life in a single age group. One rather moves smoothly from one group to the other.
- The age-group classification heuristic is ambiguous since it can be of either uniform or variable distribution.

The age groups definition problem was tackled using the AIB method. The AIB provides us with a distance unit that splits the age spectrum into the most natural groups. Unlike the ordinary AIB, in which features (x) were merged to form feature groups, in the "Age Group Definition" case, the labels (y) were merged to form age groups.

A set of experiments was conducted over the Fisher II corpus in order to find the most natural groupings. The corpus is split to 90 sub-groups, spanning from ages 20 to 64 for both genders (45 years $\times$ 2 genders). For each sub-group, a word-level feature vector is estimated. The sub-groups are then hierarchically clustered to form larger groups. Four experiments were conducted, each on a different word-level feature type:

- Word unigram vector based on the real transcription.
- Word-pairs unigram vector based on the real transcription.
- Word unigram vector based on the LVCSR-based transcription.
- Word-pairs unigram vector on the LVCSR-based transcription.

The results of the procedure are shown in Figure 2 in a dendrogram format. Since the AIB is a bottom-up approach, it is interesting to see that the two main clusters that were formed are the two genders. The four upper most clusters split each gender into two parts, a smaller part that consists mainly of the younger ages and a larger part that consists mainly of the older ages. The observation that younger age-groups consist of shorter age ranges led to the following non-linear age-group division: below 25 years old, between 25 and 40 years old, between 41 and 55 years old and above 55 years old.

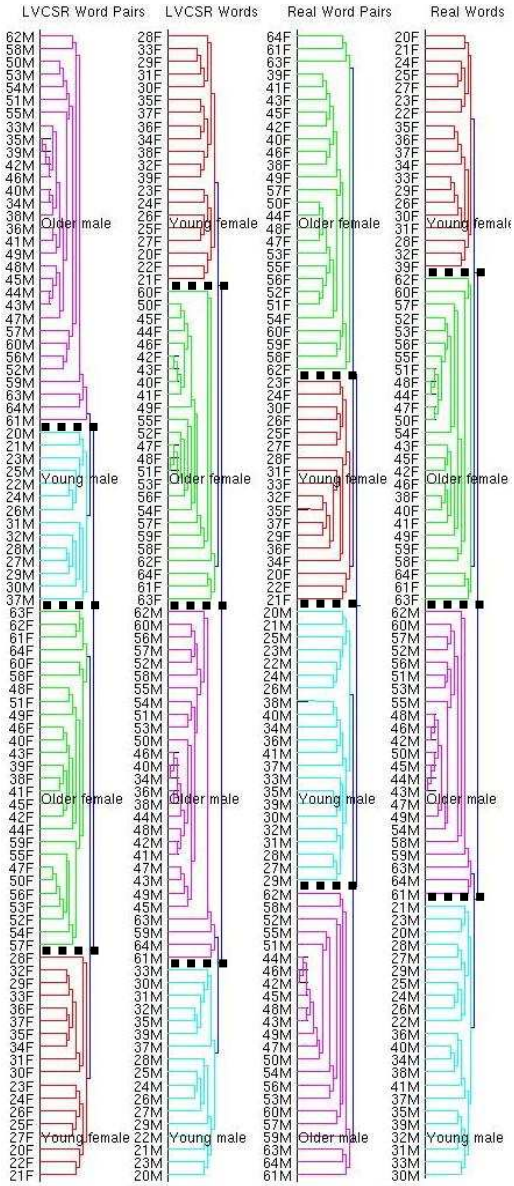

Figure 2: *Dendrogram of the age and gender groups.*

In Figure 3, the clustering process can be viewed from the mutual information aspect. For each number of clusters the mutual information between the words and the clusters is given (Eq. 1). It can be observed that significant age

clustering can be done with only a little loss of information. The decline is most significant during the last four merges.

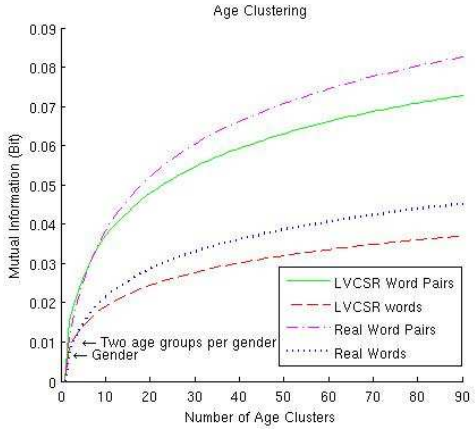

Figure 3: *Mutual information between different word level features and the age-gender groups.*

4.3. Age Group Verification Results

The recognition performances of the different systems are shown in Figure 4. The performances are measured in average equal error rates (EER) for the different age groups. The baseline system (one that does not use the AIB clustering mechanism) performance is the two solid lines in the figure. It can be seen that they are sensitive to the number of N-grams used. The decline in performance from a 1000 words unigram system to a 50 words unigram system is almost 10% average EER.

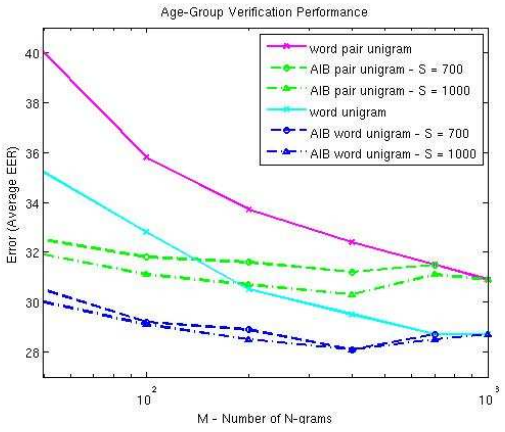

Figure 4: *Recognition performance (Average Equal Error Rate). S is the size of the original model from which the AIB started its clustering mechanism.*

Two sets of experiments were conducted using the AIB approach. Word unigrams features were used for the first set of experiments and word-pair unigrams were used in the second set. Each set was divided into three subsets: two subsets of experiments using the AIB approach and a subset of baseline experiments. The first subset was derived from the 1000 unigram system (represented in the graph by the dash-dot lines) and the second subset was derived from the 700 unigram system (represented by the dashed lines). The number of N-grams was reduced using the AIB algorithm.

It is shown that although the number of clusters was reduced, the system performance was improved. This might indicate that the information discarded was mainly irrelevant information. Where a system of 1000 word-unigrams yields a 28.7% average EER, a system comprised of 400 word-unigram clusters outperforms it with a 28.1% average EER. Another advantage that was observed is the drastic decline in model size without any loss of performance. A system of 200 word-unigram clusters and one with 1000 word-unigrams yield similar results although one is a fifth of the size of the other.

5. Conclusions

In this work, it was demonstrated that by applying the AIB method the N-gram model size can be reduced with only a small loss of relevant information. Two of the several advantages that this smaller model provides were shown. One advantage is the significant reduction in the model size. In some experiments the dictionary size was reduced from 1000 words to 50 clusters with a loss of less than 2% in average EER compared to a loss of about 10% using a 50-word model. Another advantage is the model's ability to improve verification performance. In some of the AIB experiments there was up to 0.6% improvement in the average EER compared to the baseline system.

The AIB provides a distance measurement between age groups, where each group represents a single year in the age spectrum, according to the age-group word distribution. By having the full matrix of the distances between all the age groups a hierarchal clustering was generated. Remarkably, the unsupervised clustering formed four natural age-gender groups: young-male, young-female, older-male and older-female.

6. References

[1] Doddington, G., "Speaker Recognition Based on Idiolect Differences between Speakers", in Proc. of Interspeech, 2001.

[2] Hecht, R. M., Hezroni, O., Manna, A., Aloni-Lavi, R., Dobry, G., Alfandary, A. and Zigel, Y., "Age Verification Using a Hybrid Speech Processing Approach", in Proc. of Interspeech, 2009.

[3] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A. and Leek, T. R., "High-Level Speaker Verification with Support Vector Machine ", in Proc. of ICASSP, 2004.

[4] Tur, G., Shriberg, E., Stolcke, A. and Kajarekar, S., "Duration and Pronunciation Conditioned Lexical Modeling for Speaker Verification", in Proc. of Interspeech, 2007.

[5] Shriberg, E., "High-Level Features in Speaker Recognition", Speaker Classification I, pp 241-259, Springer, 2007.

[6] Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J. G. and Little, B., "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications", in Proc. of ICASSP, 2007.

[7] Tishby, N., Pereira, F. and Bialek, W., "The Information Bottleneck Method", The 37th annual Allerton conference on communication, control and computing, 1999.

[8] Hecht, R. M. and Tishby, N., "Extraction of Relevant Speech Features Using the Information Bottleneck Method", in Proc. of Interspeech, 2005.

[9] Slonim, N. and Tishby, N., "Document Clustering Using Word Clusters Via the Information Bottleneck Method ", in Proc. of SIGIR, 2000.

[10] Schneidman, E., Slonim, N., Tishby, N., de Ruyter Van Steveninck, R. and Bialek, W., "Analyzing neural codes using the information bottleneck method", in Proc. of Advances in Neural Information Processing System (NIPS-13), 2002.

[11] Slonim, N. and Tishby, N., "Agglomerative Bottleneck Method", in Proc. of Advances in Neural Information Processing System (NIPS-12), 2000.

[12] Cieri, C., et al. "Fisher English Training" Linguistic Data Consortium, Philadelphia, 2005.