Learning Essential Speaker Sub-space Using Hetero-Associative Neural Networks for Speaker Clustering

Shajith Ikbal, Karthik Visweswariah

IBM India Research Lab, Bangalore, India

{shajmoha,v-karthik}@in.ibm.com

Abstract

In this paper, we present a novel approach to speaker clustering involving the use of hetero-associative neural network (HANN) to compute very low dimensional speaker discriminatory features (in our case 1-dimensional) in a data-driven manner. A HANN trained to map input feature space onto speaker labels through a bottle-neck hidden layer is expected to learn very low dimensional feature subspace essentially containing speaker information. The lower dimensional features are further used in a simple k-means clustering algorithm to obtain speaker segmentation. Evaluation of this approach on a database of real-life conversational speech from call-centers show that clustering performance achieved is similar to that of the state-of-the-art systems, although our approach uses just 1-dimensional features. Augmenting these features with the traditional mel-frequency cepstral coefficients (MFCC) features in the state-of-the-art system resulted in improved clustering performance.

Index Terms: hetero-associative neural network, feature extraction, speaker clustering.

1. Introduction

The aim of speaker clustering is to identify segments belonging to a single speaker in a speech utterance recorded during conversation between multiple speakers. Identifying such homogeneous speaker segments is important for tasks such as automatic speech recognition (ASR), information retrieval, and audio indexing. For example, knowing 'who spoke when' is extremely useful in improving the recognition accuracy of the ASR systems through efficient speaker adaptation procedures such as vocal tract length normalization (VTLN) [1], maximum likelihood linear regression (MLLR) [2], and feature-space MLLR (FMLLR) [3, 4]. Similarly, knowing 'who spoke what' can potentially be used to improve text analytics procedures that follow ASR output in scenarios involving call-center speech analytics.

Traditional approaches to speaker clustering are either top-down or bottom-up. Top-down approaches start with a single speaker and detect and add speakers in succession, for example, using evolutive hidden Markov models (E-HMM) [5]. Bottom-up approaches, also called hierarchical or agglomerative clustering [6, 7, 8], start with a larger number of speakers and merge the corresponding segments until a stop point is reached, resulting in a desired number of distinct speakers. Typically in agglomerative clustering, the first step is to identify all the speech segments using a speech activity detection (SAD) system. These segments are initially assumed to belong to different speakers, and a merging algorithm is used to merge these segments into speaker-homogeneous segments. The merging procedure typically involve extraction of feature vectors from those segments to model them using Gaussian mixture models (GMM) and then to decide about merging based on value of a pre-selected criterion such as overall data likelihood [9]. Typical features used for clustering are the traditional features, as used in ASR systems, such as mel-frequency cepstral coefficients (MFCC), assuming that the speaker information present in these features would lead to a good clustering performance. However, the presence of other information, especially the spoken text information, leads to large undesired variability resulting in a need for efforts to handle them.

In this paper, we present a hetero-associative neural network (HANN) [10] based approach to extract features that hopefully would contain only relevant speaker information, resulting in a simple procedure for merging step. In fact, as will be shown in the later sections of the paper, we extract 1-dimensional features and use a very simple clustering algorithm to achieve performance similar to that of the current state-of-the-art systems. HANN is able to achieve this by learning a feature extraction procedure (suitable for speaker discrimination task) through a tighter integration of feature extractor and classifier during training. In fact, such method of extracting features using HANN is not limited to only the speaker clustering problem, but could be applied in general to any pattern classification problem.

In section 2, we explain the procedure of extracting very-low dimensional class-discriminatory features using HANN, including the specifications of how we extract 1-dimensional features as used in this paper for speaker clustering task. In section 3, we explain a simple algorithm based on k-means clustering used to perform speaker clustering utilizing those 1-dimensional features. In section 4, we explain the experimental setup used to evaluate our approach, including the database and a state-of-the-art system used to establish a baseline. In section 5, we present and discuss the experimental results. In section 6, we conclude and discuss the potential future extensions of this work.

2. Feature extraction using HANN

Hetero-associative neural network (HANN) is basically a feed-forward neural network [10] with structural constraints, used to perform input feature to class-label mapping. The typical architecture of a HANN is given in Figure 1. It has an input layer, an output layer, and one or more hidden layers. Input layer has number of neuronal units equal to the input feature dimension. Number of units in the output layer is equal to the number of distinct class labels assigned to input features. Hidden layers have one or more units. One of the hidden-layers, called the bottle-neck layer, has number of units lesser than any other layer in the network. The presence of bottle-neck layer makes HANN an interesting candidate for extracting class-discriminatory fea-

10.21437/Interspeech.2008-5

tures in a data-driven manner.

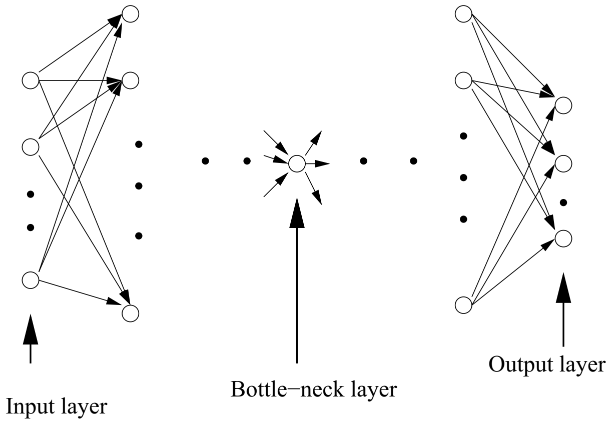

Figure 1: *Architecture of a hetero-associative neural network (HANN).*

HANN, as shown in the figure 1 can be divided into two parts: 1) part of the network from the input layer to the bottle-neck layer, and 2) part of the network from bottle-neck layer to the output layer. The first part of the network can be viewed as a feature extractor as it effectively transforms the input space onto a lower dimensional space. Similarly, the second part of the network can be viewed as a classifier as it effectively performs classification of the features obtained out of the first part of the network. Thus, HANN is basically a framework where the feature extractor and a classifier are integrated tightly together.

Given a training set of input features and the corresponding class labels, training a HANN would basically make it learn the feature extractor and the classifier in conjunction with each other. At the end of the training, feature extractor learned in the network is expected to generate features that mainly contain class discriminatory information. This is because projection onto a lower dimensional space would discard unwanted information from the input space. However, it would still retain the relevant class discriminatory information to the maximum possible extent as otherwise it would result in poor classification performance. A counter-part of HANN called auto-associative neural network (AANN), which tries to map the input onto itself through a bottle-neck layer, has been analysed extensively in the literature to show that its feature extractor part projects the input space onto a maximum variance non-linear subspace, thus performing a non-linear equivalent of principal component analysis (PCA) [11, 12]. Similarly, feature extraction part of HANN is expected to project the input space onto a maximum classification information space, thus performing a non-linear equivalent of linear discriminant analysis [13] (LDA).

2.1. Speaker-discriminatory features

Features used for speaker clustering experiments are 1-dimensional, computed using a 5-layer HANN trained with ground-truth speaker labels of about 2 minutes of speech from a single conversation involving two speakers. We have chosen to compute 1-dimensional features because if these features can be shown to achieve good clustering accuracy that would validate our claim that HANN learns essential classification information subspace. Also we have chosen to use a five layer network because that is the minimum number of layers required to perform nonlinear transformations at the feature extractor and classifier parts of the network. Input to the HANN feature extractor is standard 39 dimensional mel-frequency cepstral coefficient (MFCC) features, composed of 13 static coefficients, 13 delta coefficients, and 13 acceleration coefficients, computed from frames of 25 msec length and 10 msec shift. Network structure used is 39L-10N-1N-10N-2S, where each number denotes the number of units used in the corresponding layer and the symbols L, N, and S denote the type of output functions used in the neuronal units namely linear, non-linear sigmoid, and non-linear softmax respectively. Input layer has 39 linear units corresponding to the dimension of the input feature vector used. Output layer has 2 units, corresponding to the 2 target speaker labels. Output function used for these two output units is softmax because the network is trained in a classification mode. Bottle-neck layer, from which the 1-dimensional features are extracted, has one unit with sigmoidal output function. The sigmoidal output function restricts the value of the 1-dimensional feature to be between 0 to 1. Although, in the work for this paper, we have chosen to use 10 units each at the first and third hidden layers, varying these numbers is expected to not make much difference provided the network is trained well, unless these numbers are very small.

For the HANN training, frames from only the high energy regions of speech are used. This is because, the low energy frames, especially silence frames, are assumed to not have speaker specific information and thus including them would only distract the training. Cross-entropy criterion is used during the back-propagation training of HANN [10]. Convergence of the classification accuracy during network training is shown in Figure 2. As can be seen from the figure, classification accuracy achieved on the training data at the end of the training is about 90%. Interestingly, 1-dimensional feature at the output of the bottle-neck layer is able achieve this accuracy. We hope this is possible only when the feature subspace captured by the HANN is along the speaker discriminatory subspace. Thus speaker related variation in speech signal is expected to get reflected as variations in the values of the 1-dimensional feature.

Given a test speech utterance, 1-dimensional features to be used in segmentation algorithm are computed in a two-pass procedure as follows: Using the initial HANN (which in our case is trained with 2-minutes of speech as mentioned above), 1-dimensional features corresponding to high energy regions are computed. These features are then mapped to intermediate 2 speaker labels using a Viterbi alignment algorithm with minimal durational smoothing constraints. The value of minimum duration used is 1 minute. These intermediate speaker labels are then used to retrain HANN, with aim to tune HANN towards the speaker differences observed in the test speech. The final 1-dimensional features are computed at the bottle-neck layer of the retrained HANN.

3. Speaker clustering

Figure 3 shows 1-dimensional features computed using HANN for a segment of conversational speech involving two speakers. It also shows the ground-truth of speaker labels for the same segment. (Note that in the figure, 1-dimensional features are assigned zero values for low energy frames.) As can be seen from the figure, the 1-dimensional features are able to capture the speaker variation. In this particular case, broadly there are two clusters corresponding to two speakers present in the conversation. The presence of such distinctive speaker-specific clusters makes it possible to employ a simple clustering algorithm to

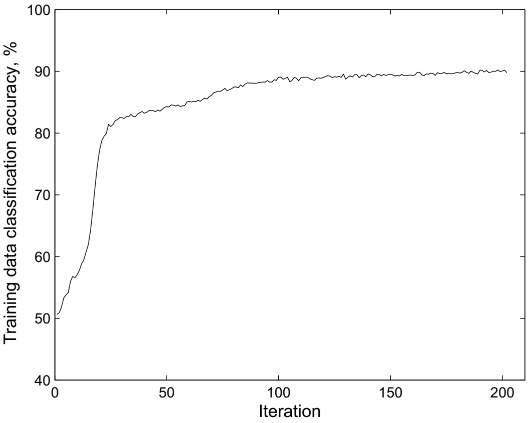

Figure 2: *Convergence of classification accuracy on training data during HANN training.*

find the speaker segments.

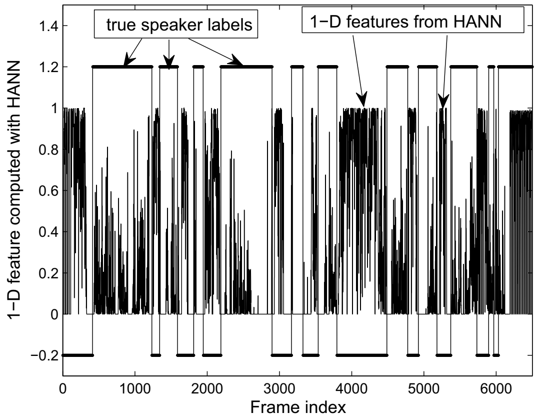

Figure 3: *1-dimensional features computed using HANN for a sample speech segment.*

The clustering algorithm to find speaker segments using 1-dimensional features is as follows: First the speech utterance is segmented into speech and non-speech using a robust adaptive algorithm [14, 15] based on just the energy, which adapts to background noise levels and energy levels in the utterance. The output of the speech/non-speech segmenter is several contiguous segments of speech. We assume that each of these segments come from a single speaker, i.e., speaker change is not possible within a segment. These segments are then clustered into k clusters (two clusters for the example shown in Figure 3) using a k-means clustering algorithm, where the absolute values of differences between the cluster mean values and the 1-dimensional features are used as the criterion. To decide about the cluster to which a particular segment is assigned, differences computed using all the 1-dimensional features in the segment are considered together.

4. Experimental setup

The database used to evaluate the proposed speaker clustering approach consists of recordings from call-center conversations, each involving two speakers, the agents and the customers. Thus the number of distinct clusters to be found out from 1-dimensional features using k-means algorithm, as explained in section 3, is fixed as 2. We have taken 100 recordings from rental car booking related conversations. Each of the conversation is about 2-3 minutes long, altogether resulting in approximately 4 hours of speech data. The sampling frequency of the recordings is 8 KHz. The recordings used are harder to deal with because they are collected in a real-life scenario with various kinds of noise and variabilities introduced by customer's location and mood such as agitation, frustration, pleasure, and satisfaction. To get an idea of the difficulty of the data in hand, best word error rate achieved with such data while performing a speech recognition experiment is about 40%.

4.1. Baseline system

In this section, we describe a baseline speaker clustering system used in this paper, that is able to achieve state-of-the-art clustering performance [16, 17] in the NIST speaker diarization evaluation. The first step is to segment into speech and non-speech, using speech/non-speech segmentation algorithm as explained in section 3 [14, 15]. The part of the data marked as non-speech is then ignored in the rest of the processing, resulting in several contiguous segments of speech. We assume initially that each of these chunks of speech comes from a single speaker. We then cluster these segments into k clusters using k-means clustering with data likelihood as the criterion and modeling each segment cluster with a single full covariance Gaussian. Finally we merge these k clusters into two clusters (for our data set we know apriori that each call contains two speakers). When merging, we merge at each step those two clusters that give the smallest loss in likelihood when merged into one cluster. MFCC features are used during clustering.

5. Results and discussion

First two lines in Table 1 gives speaker clustering accuracy achieved with our baseline approach and the proposed HANN based approach. Accuracy is computed as a percentage of time during which computed speaker labels agree with the ground-truth speaker labels. As can be seen, proposed approach is able to achieve accuracy similar to that of the baseline approach, in fact 0.5% better. This is interesting given the fact that proposed approach is able to achieve such accuracy with just 1-dimensional features, which in fact validates our claim that HANN is able to extract features that dominantly contain speaker specific information. Such 1-dimensional features and a simple k-means clustering algorithm is able to achieve performance similar to that of the state-of-the-art system.

In the next experiment, we have augmented the MFCC feature with the 1-dimensional feature computed using HANN, and used the combined feature in the baseline system. Third line in the Table 1 gives the clustering accuracy achieved, an improvement of 1.4% absolute. This clearly illustrate the effectiveness of the 1-dimensional feature.

6. Conclusions and future work

In this paper, we have presented a novel approach to speaker clustering using hetero-associative neural network (HANN).

Table 1: *Speaker clustering accuracies achieved with baseline system, proposed system, and a combination of them.*

approach	accuracy, %
baseline	77.4
HANN based	77.9
baseline+HANN	78.8

This approach uses HANN to learn very low dimensional (1-dimensional) speaker discriminatory features and use them in a simple k-means clustering algorithm to find out speaker segments. These 1-dimensional features are able to achieve clustering performance similar to that of the state-of-the-art system. In addition, augmenting these features with the standard MFCC features in a state-art-the-art speaker clustering system resulted in an improved performance, clearly demonstrating the effectiveness these 1-dimensional features. This in fact, validates our claim that HANN is able to learn class-discriminatory subspace.

In the work for this paper, 1-dimensional speaker discriminatory features are derived from standard MFCC features by feeding them at the input of the HANN. However, MFCC features are originally developed for speech recognition task and keeping that in mind MFCC feature extraction procedure goes through steps to smooth out irrelevant variabilities, including speaker variability. An interesting trial to make in the future is to learn the lower dimensional speaker discriminatory features directly from the speech signal or its spectral representation. Another interesting trial would be to extract long-term speaker related variabilities by performing feature extraction on a longer window of speech. Apart from this, for this paper, HANN was trained using ground-truth speaker labels of just 2 minutes length of speech. We can expect an improved performance by training HANN with more data. However this requires a careful pre-selection of utterances to coherently contribute to the learning of discriminatory feature subspace. Also, in this paper, the proposed approach is illustrated on a database of conversational speech involving 2 speakers. An interesting extension would be to evaluate using speech involving more than two speakers in the same conversation.

Finally, but most importantly, in this paper, HANN is shown to be a promising data-driven discriminatory feature extractor just for the case of speaker clustering task. However, this approach of discriminatory feature extraction using HANN is in general applicable to any pattern recognition problem, specifically would be interesting to try this approach to extract class discriminatory features for speech recognition task.

7. References

[1] Lee L., and Rose R. C.,"Speaker normalization using efficient frequency warping procedures", in Proc. of ICASSP, Atlanta, Georgia, USA., vol.1, pp. 353-356, 1996.

[2] Legetter C. J., and Woodland P. C.,"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", in Computer Speech and Language, pp. 171-185, 1995.

[3] Gales M. J. F.,"Maximum likelihood linear transformations for HMM-based speech recognition", in Technical Report CUED/F-INFENG, Cambridge University Engineering Dept., 1997.

[4] Vaibhava Goel, Karthik Visweswariah, and Ramesh Gopinath,"Rapid adaptation with linear combinations of rank-one matrices", in Proc. of ICASSP, 2002.

[5] Fredouille C., and Senay S.,"Technical improvements of the E-HMM based speaker diarization system for meeting records", in S. Renals, S. Bengio, and J. G. Fiscus [Ed], Machine Learning for Multimodel Interaction, Springer, vol. LNCS 4299, pp. 359-370, 2006.

[6] Reynolds D. A., and Torres-Carrasquillo,"Approaches and applications of audio diarization", in Proc. of ICASSP, vol. 5, pp. 953-956, Philadelphia, PA, 2005.

[7] Zhu X., Barras C., Meignier S., and Gauvain J.-L.,"Combining speaker identification and BIC for speaker diarization", in Proc. of Interspeech, pp. 2441-2444, Lisbon, Portugal, 2005.

[8] Anguera X., Wooters C., and Hernando J.,"Purity algorithms for speaker diarization of meeting data", in Proc. of ICASSP, vol. 1, pp. 1025-1028, Toulouse, France 2006.

[9] Ajmera J., and Wooters C.,"A robust speaker clustering algorithm", in Proc. of ASRU, St. Thomas, US Virgin Islands, 2003.

[10] Bishop, C. M.,"Neural networks for pattern recognition", Oxford: Clarendon press, 1995.

[11] Kramer M. A.,"Nonlinear principal component analysis using autoassociative neural networks", in AIChe Journal, vol. 37, pp. 233-243, 1991.

[12] Shajith Ikbal, Hemant Misra, and Bayya Yegnanarayana,"Analysis of autoassociative mapping neural networks", in Proc. of IJCNN, Washington, Jul. 1999.

[13] Duda R. O., and Hart P. E.,"Pattern classification and scene analysis", A Wiley interscience publication.

[14] Etienne Marcheret, Karthik Visweswariah, and Gerasimos Potamianos,"Speech activity detection fusing acoustic phonetic and energy features", in Proc. of Interspeech, Lisbon, Portugal, 2005.

[15] Monkowski M.,"Automatic gain control in speech recognition system", U.S. Patent, US6314396.

[16] Jing Huang, Etienne Marcheret, Karthik Visweswariah, and Gerasimos Potamianos,"The IBM RT07 evaluation systems for speaker diarization on lecture meetings", in Proc. of NIST RT07 Evaluation Workshop, Washington, 2007.

[17] Jing Huang, Etienne Marcheret, Karthik Visweswariah, Vit Libal, and Gerasimos Potamianos,"Detection, diarization, and transcription of far-field lecture speech", in Proc. of Interspeech, Antwerp, Belgium, 2007.