MUTUAL INFORMATION BASED CHANNEL SELECTION FOR SPEAKER DIARIZATION OF MEETINGS DATA

Deepu Vijayasenan[1,2], Fabio Valente[1], Hervé Bourlard[1,2] *

[1]Idiap Research Institute, 1920, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), 1015, Lausanne,Switzerland
{deepu.vijayasenan,fabio.valente, herve.bourlard}@idiap.ch

ABSTRACT

In the meeting case scenario, audio is often recorded using Multiple Distance Microphones (MDM) in a non-intrusive manner. Typically a beamforming is performed in order to obtain a single enhanced signal out of the multiple channels. This paper investigates the use of mutual information for selecting the channel subset that produces the lowest error in a diarization system. Conventional systems perform channel selection on the basis of signal properties such as SNR, cross correlation. In this paper, we propose the use of a mutual information measure that is directly related to the objective function of the diarization system. The proposed algorithms are evaluated on the NIST RT 06 eval dataset. Channel selection improves the speaker error by 1.1% absolute (6.5% relative) w.r.t. the use of all channels.

Index Terms— Speaker diarization, Information Bottleneck clustering, Channel selection, Mutual information

1. INTRODUCTION

Speaker diarization determines "*who spoke when*" in a given audio recording. This involves finding the number of speakers and identification of speech segments of each speaker in an unsupervised manner.

In the meeting case scenario, data acquisition is done in a non-intrusive manner using a microphone array often referred as Multiple Distant Microphone (MDM). Conventional diarization systems use a single data stream, thus the signals from multiple channels are used to produce a single enhanced signal typically by a beam-forming algorithm. For a review of the use of beam-forming algorithms in speaker diarization see [1].

The beam-forming algorithm used in [2] selects a reference channel based on the average cross-correlation and then performs a delay-and-sum combination. Delays are computed with respect to the reference channel.

Data used in the NIST Rich Transcription evaluation consists of meetings recorded in several sites. These meetings represent a very heterogeneous data set because of varying number, topology and quality of microphones in the array. In order to increase the robustness of the beam-forming algorithm to the different conditions, several channel weighting and channel selection algorithms have been tested in [1] (chapter 5). Typically channel weighting and channel selection is performed on the basis of Signal to Noise Ratio (SNR) or average cross correlation between channels. However, SNR and cross-correlations are not directly related to the diarization performance of the system.

In this paper, we consider the problem of channel selection from a heterogeneous collection of microphones using a measure directly related to the diarization system. Channel selection is related to the fact that sometimes the quality of a channel can be so low that its use would in any case degrade the performances. In particular, the following issues are addressed:

1 How to select the channel that provides the lowest diarization error out of the available channels.

2 How to select the subset of channels that provides the lowest diarization error.

Previously we have proposed a system [3] based on the Information Bottleneck (IB) principle [4] which is inspired from Rate-Distortion theory. The speaker diarization aims at finding the clustering that minimize the loss in mutual information between the initial uniform segmentation and the final clustering. Instead of using measures related to the signal itself (SNR or cross-correlations), we investigate the use of the mutual information as measure for assessing the quality of a channel or subset of channels. This is based on the assumption that highest mutual information will produce better clustering. The paper is organized as follows. Section 2 summarizes the IB principle and Section 3 presents the speaker diarization system using the IB framework. The channel selection scheme for selection of the best channel and a channel subset is detailed in Section 4. Section 5 describes all the experiments performed as well as the baseline system. Section 6 concludes the paper.

2. INFORMATION BOTTLENECK PRINCIPLE

Let X, be a set of elements to cluster into a set of C clusters, for instance a set of speech segments. Let Y be a set of variables of interest associated with X such that $\forall x \epsilon X$ and $\forall y \epsilon Y$ the conditional distribution $p(y|x)$ is available. In speaker diarization, we use the components of a background GMM as the relevance variables. Clusters C can be interpreted as a compression (bottleneck) of initial data set X in which information that X contains about Y is passed through the bottleneck C. The Information Bottleneck (IB) principle states that the clustering C should preserve as much information as possible from the original data set X w.r.t. relevance variables Y.

IB method [4] is inspired from Rate-Distortion theory which states the best representation C of data X minimizes the mutual information $I(X, C)$, i.e. the distortion and preserves as much information as possible about Y (maximizing $I(C, Y)$). Thus the

*This work was supported by the European Union under the integrated projects AMIDA, Augmented Multi-party Interaction with Distance Access, contract number IST-033812, as well as KERSEQ project under the Indo Swiss Joint Research Program (ISJRP) financed by the Swiss National Science Foundation. This project is pursued in collaboration with EPFL under contract number IT02.

IB objective function can be formulated as minimization of the Lagrangian,

$$I(X,C) - \beta I(C,Y) \quad (1)$$

where β is the trade-off between the amount of information $I(C,Y)$ to be preserved and the compression of the initial representation $I(C,X)$. Function (1) must be optimized w.r.t. the stochastic mapping $p(C|X)$. Expressions for $I(X,C)$ and $I(C,Y)$ can be developed as $I(X,C) = \sum_{x,c} p(x)p(c|x)log\frac{p(c|x)}{p(c)}$ and $I(C,Y) = \sum_{y,c} p(c)p(y|c)log\frac{p(y|c)}{p(y)}$ This leads to a set of self-consistent equations as shown in [5] which can be solved to obtain the cluster representation.

The limit $\beta \to \infty$ induces a hard partition of the input space i.e. the probabilistic map $p(c|x)$, takes values of 0 and 1 only. This is equivalent to minimizing only the information loss in the clustering i.e. $I(C,Y)$. Different algorithms have been proposed in literature for minimizing the IB objective function. One common approach is the agglomerative Information Bottleneck(aIB).

2.1. Agglomerative Information Bottleneck

The agglomerative Information Bottleneck (aIB) is a greedy approach to minimize the objective function of equation (1). The initialization consists of the trivial clustering of $|X|$ clusters; i.e. each data point treated as a separate cluster. Subsequently the clusters are merged iteratively such that after each step the loss of mutual information w.r.t the relevant variables Y is minimum.

The loss of mutual information δI_y obtained by merging two clusters x_i and x_j is given by Jensen-Shannon divergence between $p(Y|x_i)$ and $p(Y|x_j)$ (see [5]). In case of discrete probabilities, this divergence is straightforward to compute. The information preserved in each step decreases monotonically. Details about implementation of aIB algorithm can be found in [5] and will not be further discussed here. The optimal number of clusters is selected by thresholding the Normalized Mutual Information, $NMI = \frac{I(C,Y)}{I(X,Y)}$. Details of this method are described in [3].

3. SPEAKER DIARIZATION ALGORITHM

We summarize here the speaker diarization algorithm described in detail in [3]. The clustering steps are described below.

1. Acoustic feature extraction from the audio file.
2. Speech/non-speech segmentation and rejection of non-speech frames and uniform segmentation of speech frames in chunks of fixed size D = 2.5s i.e., definition of set X.
4. Estimation of GMM model with shared diagonal covariance matrix for each segment i.e., definition of set Y.
5. Estimation of conditional probability $p(y|x)$.
6. IB Clustering and model selection using NMI as described in section 2.1.
7. Viterbi realignment using conventional HMM/GMM system estimated from previous segmentation.

This clustering relies on the purity of initial segments X which are arbitrarily obtained by uniform segmentation. If the length of the segment D is small enough segments may be considered as generated by a single speaker. Although this hypothesis can be true in case of Broadcast News audio data, in case of conversational speech with fast speaker change rate and overlapping speech (like in meetings data), initial segments may contain speech from several speakers. Thus Viterbi re-alignment is performed in order to refine the segment boundaries.

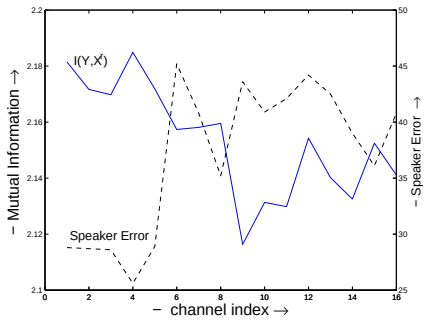


Fig. 1. Correlation between Mutual information and DER for the meeting EDI_20050216-1051 for different channels. Higher mutual information corresponds to lower diarization errors

4. MUTUAL INFORMATION CHANNEL SELECTION

Let us consider N different acoustic feature streams (MFCC coefficients). Let $X_i, i = 1,\ldots,N$, denote the speech chunks (input variables for clustering) extracted from the acoustic feature streams. Consider the objective function of equation (1) for each stream:

$$\mathcal{F}_i = I(X_i,C_i) - \beta I(C_i,Y) \quad (2)$$

When the aIB clustering starts ($C_i = X_i$) the value of the objective function is given by:

$$H(X_i) - \beta I(X_i,Y) \quad (3)$$

since $I(X_i,X_i) = H(X_i)$. All X_i are derived using the same segmentation, thus the random variables have identical distribution $p(x_i)$, and hence the same $H(X_i)$. This implies the feature stream that has the maximum $I(X_i,Y)$ minimizes the objective function. In other words minimizing the objective function (1) at the beginning of the clustering, it is equivalent to maximizing the mutual information $I(X_i,Y)$.

Intuitively, the quality of the clustering will depend on how informative the relevance variables Y are on the speech segments X_i i.e. on $I(Y,X_i)$. We can expect that higher initial $I(Y,X_i)$ will produce better clustering. Based on similar considerations, mutual information of relevance variables with X was used for feature selection in the problem of text processing ([6] Chapter 4). As an example, Figure 1 shows the correlation of diarization error and $I(Y,X_i)$ for various channels of one meeting in RT 06 eval dataset.

In the following we consider two different problems: the selection of the single channel and the selection of the channel subset that will produce the lowest Diarization Error Rate. In the first case, obtaining $I(Y,X_i)$ involves the extraction of separate acoustic features for each channel. In the second case, obtaining $I(Y,X_i)$ involves the selection of a subset of channels, the beamforming of this subset and the extraction of the acoustic features. This is equivalent to the approach referred as channel elimination in [1] although based on a different criterion. Let us consider separately the two scenarios.

4.1. Single Channel Selection

Let us consider a set of N channels and N different acoustic feature streams (MFCC coefficients). Here we aim to select the best channel that provides the minimum diarization error. Let S_i^c, $i = 1,\ldots,N$ be the MFCC features extracted from each channel (The superscript

denotes the features correspond to individual channels). Each acoustic feature stream S_i^c is segmented in chunks of fixed size D = 2.5s. Let X_i^c denotes the speech chunks that correspond to the acoustic feature stream S_i^c. In order to compare the mutual information as obtained from different X_i^c, the relevance variable set Y must be kept fixed. We define Y as the components of a GMM background model as obtained from the beamforming of *all* the available channels.

Thus, the best channel is chosen as $i^\star = \arg\max_i I(Y, X_i^c)$

4.2. Multiple Channel Selection

In case of channel subset selection, the goal is to determine a subset of the available channels that once beamformed, produces the lowest diarization error. The brute force solution would involve the beamforming of all possible subsets. In case of N channels, the number of possible non-empty subsets to be considered is equal to $2^N - 1$. This value can be prohibitively high like in case of EDI meetings recorded with 16 channels.

Instead, we adopt a greedy approach that uses the channel ranking as discussed in section 4.1. The algorithm is summarized as follows:

1 Sort the N single channels according to the value of $I(Y, X_i^c)$ obtained as in section 4.1.

2 Consider N possible subsets obtained from the sorted list such that the first subset contains the top channel, the second subset contains the top two channels and so on. This is equivalent to having an $n-$best list of channels and uses N possible subsets instead of $2^N - 1$.

3 Perform beamforming on these subsets.

4 Extract MFCC coefficients for each of the N beamformed signals. Let S_k^b be the MFCC feature stream extracted from the *beamformed* output of top k channels.

5 Perform uniform segmentation of speech in fixed chunks to define input variables for clustering. Let X_k^b be the input variables that correspond to S_k^b.

6 Compute $I(Y, X_k^b)$ using a background GMM model. As before the GMM estimated from the beamforming of all available channels is used.

7 Select the best channel subset as $k^\star = \arg\max_k I(Y, X_k^b)$

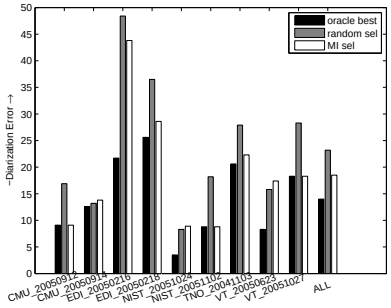
This algorithm will select the subset k that, once beamformed, will produce the highest mutual information. One advantage with the approach is that no thresholds are involved in channel selection unlike conventional methods which depends on SNR or cross correlation [2]. The method is actually a greedy approach to the channel elimination and we will experimentally verify its effectiveness.

5. EXPERIMENTS AND RESULTS

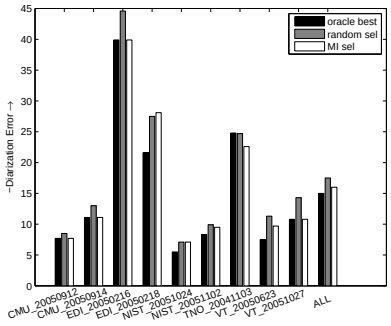
We performed all the experiments on the NIST RT06 evaluation data for "Meeting Recognition Diarization" task based on data from Multiple Distant Microphones (MDM) [7] and results are provided in terms of Diarization Error Rates (DER). DER is the sum of missed speech error, false alarm speech error and speaker error (for details on DER see [8]). Speech/non-speech (spnsp) is the sum of missed speech and false alarm speech. System parameters are tuned on the development data. We used the NIST RT05 evaluation data as the development data. Delay and sum beamforming is performed with the *BeamformIt* toolkit [9]. 19 MFCC features are then extracted from the signal.

Table 1. Speech/No speech, speaker error and DER of the baseline system. All channels are used in the beamforming

Miss	FA	spnsp	spkr err	DER
6.5	0.1	6.6	17.1	23.7



(a)



(b)

Fig. 2. Meeting-wise speaker error rates for Oracle , random, and mutual information based selection. (a) Selection of single channel (b) Selection of a subset

Speech/non-speech segmentation is obtained using a forced alignment of the reference transcripts on close talking microphone data using the AMI RT06s first pass ASR models [10]. Results are scored against manual references force aligned by an ASR system.

The results of the baseline system as discussed in Section 3 is presented in Table 1. In this approach all channels are used for beamforming. Since the we use the same speech nonspeech segmentation for all the experiments, only speaker error is reported hereafter. The sp/nsp error of all algorithms discussed in this paper will be same as presented in Table 1.

5.1. Single Channel selection

In this experiment, we tried to select a single best channel as described in Section 4.1. We performed oracle channel selection (exhaustively computing the diarization error for each channel and manually selecting the channel with highest/lowest DER) to compare the proposed method with the best and worst case scenario. We also performed random selection of channels to ensure the algorithm performance is better than chance. A channel is selected at random with uniform prior assigned for all the channels. The average speaker error is calculated from multiple trials (10 trials).

Table 2. Speaker error for different channel selection algorithms

Selection Scheme	spkr err (%)
Oracle best	14.3
Oracle worst	29.1
Random selection	23.2
Max Cross correlation	22.9
Max $I(Y, X_r)$	**18.5**

(a) Single channel selection

Selection Scheme	spkr err (%)
Oracle best	15.3
Oracle worst	21.0
Random selection	18.4
Selection using $I(Y, X_r)$	**16.0**

(b) Multiple channel selection

We also performed channel selection using maximum average cross correlation as discussed in [1]. In this method a single channel is selected as follows. The average cross correlation of each channel with respect to all other channels is computed. The channel with maximum average cross correlation with all other channels is then selected for diarization. Table 2(a) lists the results of various schemes. The proposed system performance is better than the random selection system or using cross correlation based channel selection. Note that the total speaker error is close to that of the baseline just with using one best channel. Figure 2(a) lists meeting-wise speaker errors. The proposed scheme is better than the random selection in most of the meetings.

5.2. Multiple Channel selection

In this set of experiments, we select a subset of the channels rather than single best channel. We perform the oracle best, worst and the random selection experiments as in the single channel case. Note that, only N subsets are considered in this step instead of $2^N - 1$ possible subsets. The system performs significantly better than the random subset selection (Table 2(b)). Meeting-wise speaker errors (Figure 2(b)) shows proposed scheme has better performance compared to random selection[1].

The number of channels and the speaker error of each meeting for the baseline as well as the proposed system is listed in Table 3. Only half of the input channels were selected in case of EDI meetings. Each of these meetings consists of two microphone arrays. It is observed that the microphone array that had high diarization rates has been eliminated in the channel selection process. The system outperforms the baseline system by 1%. It can be seen that the DER improves in most of the cases when the system selects a subset of the channels.

6. CONCLUSIONS

In this work we observed a correlation between the diarization error of individual channels of the MDM data and the mutual informations of the extracted features with respect to a background GMM. We proposed to use this information to select a single best channel for diarization. The results are better than randomly selecting channels by 4.7% absolute. Even with using only one channel the results are comparable (1.4% worse absolute) to the baseline system which uses beamforming of all the channels. On the other hand, when random channel selection is performed the system is significantly worst than baseline (6.1% worse absolute).

We also proposed an algorithm based on mutual information to select channel subset for beamforming. This is based on the fact that some channels have such poor performance that would not anyway help in combination with others. The algorithm performance is very close to selecting the best subset manually (0.7% worse than oracle). It uses only a subset of channels for 5 meetings and it performs 1% absolute better then the system that uses all the channels. As in the case of meetings like EDI, the algorithm eliminates the microphone array that yields poor performance. Also in case of NIST meetings only a subset of the array is used. The multichannel selection algorithm first beamforms the channels and then computes the mutual information criterion. Alternatively, individual channels can be selected iteratively such that each newly added feature brings the maximum increase in the total mutual information. Similar approaches has been explored in the context of classifier feature selection [11], and would be investigated in future.

Table 3. Total number of channels and number of channels selected for each meeting with corresponding speaker errors

	Proposed scheme		Baseline	
Meeting	spkr err	#channels	spkr err	#channels
CMU_20050912	7.7	2	7.7	2
CMU_20050914	11.1	2	11.1	2
EDI_20050216	39.9	8	46.0	16
EDI_20050218	28.1	8	29.6	16
NIST_20051024	7.1	5	9.1	7
NIST_20051102	9.5	6	9.4	7
TNO_20041103	22.6	9	22.6	10
VT_20050623	9.7	4	9.7	4
VT_20051027	10.8	3	10.8	3

7. REFERENCES

[1] Xavier Anguera, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politecnica de Catalunya, 2006.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[3] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2007, pp. 250–255.

[4] N. Tishby, F.C. Pereira, and W. Bialek, "The information bottleneck method," in *NEC Research Institute TR*, 1998.

[5] N. Slonim, N. Friedman, and N. Tishby, "Agglomerative information bottleneck," in *Proceedings of Advances in Neural Information Processing Systems*. MIT Press, 1999, pp. 617–623.

[6] Noam Slonim, *The Information Bottleneck: Theory and Applications*, Ph.D. thesis, The Hebrew University of Jerusalem, 2002.

[7] "http://www.nist.gov/speech/tests/rt/rt2006/spring/," .

[8] "http://nist.gov/speech/tests/rt/rt2004/fall/," .

[9] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," in *http://www.icsi.berkeley.edu/xanguera/BeamformIt*, 2006.

[10] Hain T. et. al., "The AMI meeting transcription system: Progress and performance," in *Proceedings of NIST RT'O6 Workshop*, 2006.

[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks, IEEE Transactions on*, vol. 5, no. 4, pp. 537–550, 1994.

[1]The oracle best in the multiple channel selection is worse than that of single channel selection. This is because the oracle is performed prior to Viterbi realignment and the DER is evaluated after Viterbi realignment