Channel and channel subband selection for speaker diarization

Ahmed, Ahmed Isam; Chiverton, John P.; Ndzi, David L.; Al-Faris, Mahmoud M.

Published in:
Computer Speech and Language

DOI:
10.1016/j.csl.2022.101367

E-pub ahead of print: 30/09/2022

Channel and Channel Subband Selection for Speaker Diarization[☆]

Ahmed Isam Ahmed[a,*], John P Chiverton[a], David L Ndzi[b], Mahmoud M Al-Faris[a]

[a]*School of Energy and Electronic Engineering, University of Portsmouth, Portsmouth, UK, PO1 3DJ*
[b]*School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley, UK, PA1 2BE*

Abstract

Speaker diarization can be considered to be one of the complex problems in speaker recognition. A reliable diarization system should be able to accurately determine the variable length utterances which a speaker contributes to multi-speaker conversations. This is a difficult problem since text-independent speaker identification and verification is yet to be improved for it to be applied reliably. While efficient speaker modelling is important for diarization, the acoustical representation of speech is the basic entity that signifies a speaker. This representation should be outstanding enough to prevent a speaker's utterances from being lost in the acoustical congestion that is imposed by the rest of the talkers.

For this purpose, it is proposed here, for the case of multiple-microphone diarization, multiple speech signals are used in the acoustic feature extraction instead of combining the signals beforehand. The reason is to make an optimal use of those signals in order to enrich the quality of the acoustical representation of the speaker. To this end, and since not all microphone signals (channels) may be desirable, two selection approaches are proposed in this work. These are, a best quality channel selection method and a novel approach for diverse channel selection. Furthermore, a novel method is proposed which retains the speech spectrum from selected least reverberated subbands of the available channels' spectrums. A new model, referred to here as Averaged Joint Gradient (AJG), is introduced for this purpose. The proposed approach reduces the Diarization Error Rate (DER) in both of the diarization systems used in the evaluations. The first system is based on binary keys and achieves a maximum relative reduction in DER of 14%. The second one is a Gaussian Mixture Model-Bayesian Information Criterion (GMM-BIC) based system which achieves a maximum relative reduction in DER of 20%.

Keywords: speaker diarization, channel selection, reverberation, acoustic beamforming.

[☆] Declarations of interest: none

[*] Corresponding author

Email addresses: `ahmed.ahmed5@myport.ac.uk` (Ahmed Isam Ahmed), `john.chiverton@port.ac.uk` (John P Chiverton), `david.ndzi@uws.ac.uk` (David L Ndzi), `mahmoud.al-faris1@myport.ac.uk` (Mahmoud M Al-Faris)

1. Introduction

Speaker diarization is the task of determining who spoke and when in an audio stream of multiple speakers. It has applications in speech and speaker indexing as well as speech-to-text transcription to name a few examples (Anguera et al., 2012). It is a challenging task that can be deemed necessary for unsupervised speaker recognition. This is because it helps in detecting the change points between speakers. Challenging computational science tasks often benefit from increasing the number and variety of sources of data, see e.g. (Zhang et al., 2020). Speaker diarization is no different and can benefit from additional audio signal sources (Pardo et al., 2007a). In general, this is not particularly difficult to achieve, as microphones are relatively inexpensive. This work focuses on the diarization of sessions that are recorded by a set of distant stationary microphones (not attached to speakers). All of the microphones are assumed to have recorded all of the speakers. However, no assumptions are made about the number of speakers or their locations. The main body of the work is developed on meeting data, but the outcome is also tested on dinner party sessions.

When a dialogue is recorded by multiple microphones, additional information becomes available which helps the diarization task in two ways. Firstly, it is used to extract spatial features such as Time Delay of Arrival Features (TDOA) (Parada et al., 2017) and Direction of Arrival Features (DOA) (Ito et al., 2018). Secondly, the microphone signals can be combined using beamforming, e.g. (Anguera et al., 2007), to produce a single signal, enhanced for acoustic feature extraction (Martínez-González et al., 2017). This is regarded as an efficient solution. However, improvements in the diarization performance can be achieved by considering different frameworks. The aim of this work is to improve the way multiple microphone signals are exploited for acoustic feature extraction. This is accomplished by extracting the acoustic features from a number of selected microphones (or microphones' sub-bands). A number of factors may vary among recordings made by multiple microphones. The selection in this work addresses the general quality, redundancy and potentially subband-dependent undesirable effects.

A number of different speaker diarization approaches such as those presented by (Sun et al., 2018; Dawalatabad et al., 2016; Madikeri et al., 2015) have adopted the aforementioned beamforming technique of Anguera et al. (2007). This entails combining all of the channels' signals into one signal using a weighted delay and sum technique.[1] Then acoustic features, usually Mel-Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein, 1980), are extracted from the resultant signal. Recently in (Tu et al., 2017), alternative solutions were proposed to overcome the imperfection of a number of beamforming techniques. One of the problems was related to the direction of arrival mismatch. One of the proposed alternatives to using one beamformed signal was to concatenate features extracted from multiple signals. The signals were obtained using beamformed combinations of subsets of microphone signals. The solution of (Tu et al., 2017) was found to improve speech recognition accuracy but it still relied on creating and extraction of features from beamformed signals.

[1] The terms channel and microphone are used interchangeably here.

The use of beamformed signals for feature extraction can be useful for diarization because the beam of the formed signal is likely to be directed toward the talking speaker. However, beamforming entails a series of operations, see (Anguera et al., 2007), after which some genuine properties of microphone signals may become lost. Alternatively, if feature extraction is applied directly on the signals, their genuine properties should propagate to the extracted features and can help in diarization, see e.g. (Meignier and Merlin, 2010). That work recommended preserving original conditions by avoiding further processes on the features, such as normalisation. Furthermore, in the work on dimensionality reduction in (Panday et al., 2018) and (Zhang et al., 2019), feature selection methods were favoured because they do not alter the original features. Nonetheless, extracting features from all available channels is computationally inefficient and some channels are less likely to be good quality (Anguera et al., 2007).

Speech signal quality measures can be divided into intrusive and non-intrusive (see Falk et al. (2010)). Intrusive measures require a reference (such as a clean signal) while non-intrusive measures do not require a reference. The Signal to Noise Ratio (SNR) parameter was used in (Pardo et al., 2007b) to identify a good quality channel to be used as a reference in delay estimation. Although estimating the SNR is non-intrusive, it may not be feasible because it is difficult to estimate accurately (Bosworth et al., 2008).

Speech recognition techniques can often be found to include the development of signal quality measures. For instance, in Distant Speech Recognition (DSR), Wolf and Nadeu (2014) introduced a number of decoder-based measures, like the variance of the speech intensity envelope, for channel selection. However, the proposed measures were demanding as they required a classification of the recognised speech then the selection is made and the recognition is repeated using the selected channels. The modulation spectrum ratio introduced in (Himawan et al., 2015) was also used for channel selection in distant speech recognition. Original speech was convoluted with different rooms' impulse responses. Then the correlation between contaminated speech and the Word Error Rate (WER) was used to predict the recognition performance. By assuming an exact knowledge of a real room impulse response, the modulation spectrum ratio was used to select the best channel. Unfortunately, the applicability of the proposed framework could be considered to be somewhat constrained by the fact that it required an exact knowledge of a room's impulse response. It also assumed the speakers to be stationary.

Cepstral distance is another signal quality measure. This intrusive measure was introduced by (Kitawaki et al., 1988) to assess the distortion presented by speech coding techniques in reference to the original speech signal. Cepstral distance is long known for its flexibility and effectiveness in different applications (Guerrero et al., 2016). It was recently used for the selection of the least distorted channel by (Flores et al., 2018) for distant speech recognition. As an intrusive measure, the use of the cepstral distance requires a reference channel which is assumed to provide a clean speech signal in some sense.

In (Flores et al., 2018), the authors proposed to compute a reference signal from the logarithm of the geometric mean of the signals of the available microphones. The geometric mean was calculated over the magnitude spectrums generated from each microphone. Unfortunately, no distinction was made between the quality of the available signals. As a result, potentially good and bad quality signals contributed equally to the reference signal computation because of the unweighted mean element of the method. It would be more robust to assign preliminary quality-based weights in such an averaging process.

Other than signal quality, Meignier and Merlin (2010) recommended using unnormalised acoustic features in the common segmentation and clustering processes of the diarization system. The aim was to preserve the background environment which can help in differentiating between speakers. It is considered here in this work that features extracted from individual signals can better reflect the underlying condition of each signal. But, given a plethora of signals, it is anticipated that groups of signals are subjected to similar conditions. This indicates the necessity for a selection method which attempts to keep only the channels that may provide different background information amongst other properties. This is addressed in Section 2.1.

Reverberation, in particular, is a deteriorating effect on the quality of speech signals that has been the focus of considerable research efforts (Kinoshita et al., 2016). Reverberated speech develops when reflected speech is delayed and overlaps with the directly propagated speech at the acquisition point. One way to tackle reverberation is to de-reverberate the speech signal or features as in (Feng et al., 2014) for speech recognition where deep autoencoders were used for this purpose. However, de-reverberation is difficult and non-reliable since it can introduce objectionable artefacts to the processed speech (Falk et al., 2010). Alternatively, in (Giri et al., 2015), a feature vector that characterises reverberation was extracted from the speech signal and input to a Deep Neural Network (DNN) in a room-aware DNN training for speech recognition. A similar concept was presented in (Oo et al., 2018) in a reverberation-aware DNN training.

Additionally, methods have been proposed to characterise the degree of reverberation, e.g (Malik and Farid, 2010) and (Falk et al., 2010). The concept of modulation transfer function (MTF) is one of the earliest approaches applied to evaluate the quality of speech transmission (against reverberation and other effects) between the speaker and the listener in an auditorium (Houtgast and Steeneken, 1985). In (Malik and Farid, 2010), reverberation is detected by estimating a decay parameter that embodies the extent of reverberation. That parameter is estimated from the speech signal using a maximum likelihood estimation. The work in (Falk et al., 2010) introduced a measure termed speech-to-reverberation modulation energy ratio for the diagnosis of de-reverberated speech to test for the feasibility of de-reverberation algorithms. In (Jiang et al., 2014), binary classification using a DNN was introduced for reverberant speech segregation. This required the extraction of binaural features of the interaural time differences and interaural level differences which were used as the main auditory features. Interestingly, reverberation is frequency dependent (Wen et al., 2008; Li et al., 2019). This means that, depending on a room's geometry, size and surface material, the degree of reverberation varies between subbands of the speech spectrum (Ismail, 2013; Zhu et al., 2020). If this variety is to be distinguished, time-domain reverberation measures, such as the one in (Malik and Farid, 2010), are not usable. Also, precise estimation of the degree of reverberation as expected from the measures presented in (Falk et al., 2010) and (Ismail, 2013) would not be necessary.

A method proposed here is designed for the purpose of assessing the degree of reverberation over a subband in comparison to the rest of the channels. As such, the subband that is least affected by reverberation is identified and used in feature extraction. For the purpose of channel selection, this work proposes the use of a potentially reliable reference signal for selecting good quality channels based on the cepstral distance. Moreover, it presents the selection of diverse channel selection whose goal is to spare channels that may provide redundant characteristics. Section 2 presents the channel selection methods proposed in this work while Section 3 presents channels' subband selection. Sections 4 and 5 report the evaluation results in binary key based and GMM-BIC based diarization, respectively. These are followed by the conclusions in Section 6.

2. Channel Selection

This section presents the methods that aim to select suitable channels to use their acoustic features in the diarization system. The goal of the first method is to find two sets of microphones that are distant from each other. The second method aims to find the highest quality channels.

2.1. Selection of Distant Microphones

The fundamental theory behind this selection method is that when many microphones are available, then those that can deliver a diversity in information are favoured. This is based on the assumption that each microphone record all of the speakers. Therefore, a desired selection method could be one that identifies the microphones with the redundant information. These should then be discarded. Selecting microphones that are distant from each other would fulfil this goal. This is because of the proposition that variability in microphones' locations can cause the recorded speech to be subject to effects of different natures due to the room impulse response. In other words, the difference in the effect of the room impulse response on the speech recorded by each microphone is emphasised if the recording microphones are not placed close to each other (Anguera et al., 2007).

Additionally, when microphones are in variable locations, the differences in the distances between each microphone and each speaker will be broadened. The effect of this could be duplicated through the energy pattern of the recorded speech signal. This in turn can help in the diarization since a particular speaker's speech would enclose different conditions from other speakers. The diversity may also, but not necessarily, increase statistical independence within the extracted features especially when Gaussian Mixture Models (GMM) with diagonal covariance matrices are deployed in the diarization system. Note that diagonal covariance matrices better describe the covariance if the variables (the features) are more statistically independent (Deco and Obradovic, 2012).

Considering the aforementioned basis, features extracted from two distant microphones would present an improvement in the performance which is actually the case as will be shown in the results. In addition, it is found that using two groups of distant microphones can present even better performance. The suitable number of selected microphones for each group will be chosen empirically.

Before beamforming became the common practice when multiple microphones' speech is available, the centrally located microphone was usually identified based on cross correlation and was considered a good signal source for the extraction of speech features (Anguera et al., 2007). A centrally located microphone is considered a 'close' microphone in the methodology of this section. The spatial distribution of other microphones is decided based on signal time delay of arrival at any of the available microphones in relation to the central one.

For this channel selection method, the delays between the signals are calculated using the cross correlation method over segments of 250 ms length and 10 ms shift. The delays are estimated over segments of the signal in order to accommodate for potential changes in speakers' locations. Assuming that microphones are stationary, the average of segments' delays is supposed to help approximate the distribution of the speakers and the microphones. For two speech segments from one of the microphones, s_i, and the reference (central) microphone, s_{ref}, the delay τ (the lag) in samples, is the one that maximises the following cross correlation function

$$\mathcal{C}_{(\tau)} = \sum_{n=1}^{N} s_i(n) s_{\text{ref}}(\tau + n), \tag{1}$$

where N is the total number of samples within the segment.

Let τ_m be the one that maximises (1). The average of the absolute delays over all of the segments of one of the microphones i and the central ref microphone can then be expressed as

$$\tilde{\tau}_{i,\text{ref}} = \frac{1}{S} \sum_{j=1}^{S} \tau_m(s_i(j), s_{\text{ref}}(j)), \tag{2}$$

where S is the total number of segments.

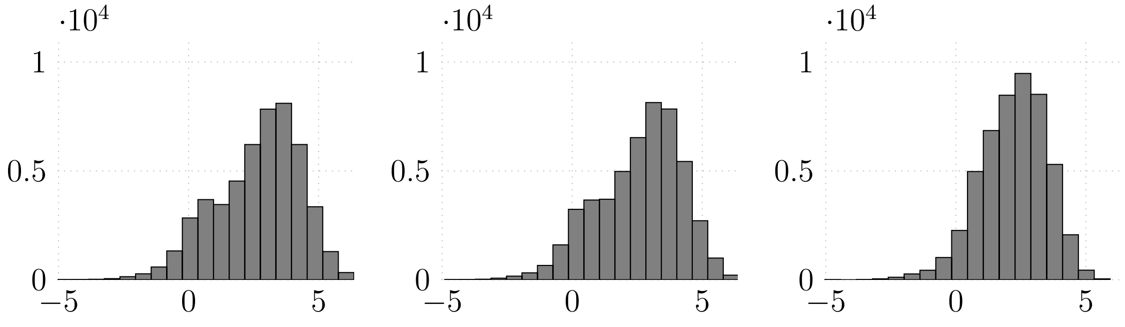
For a total number of M microphones, the farthest microphone from the central ref one is selected as the one with the highest $\tilde{\tau}_{i,\text{ref}}$

$$\arg\max_{\forall i} \quad \tilde{\tau}_{i,\text{ref}} \quad \text{for} \quad i = 1, 2, ..., M-1, \tag{3}$$

and the closest microphone to the central ref microphone is the one with the lowest $\tilde{\tau}_{j,\text{ref}}$

$$\arg\min_{\forall j} \quad \tilde{\tau}_{i,\text{ref}} \quad \text{for} \quad j = 1, 2, ..., M-1. \tag{4}$$

The first order coefficient of MFCC features reflects the distribution of speech spectral energy between low and high frequencies of speech. It can be used to demonstrate the diversity between distant microphones. Fig. 1 shows the distribution of this coefficient extracted from signals of the central, the nearest and the farthest microphones of the IS1001a AMI meeting corpus (Carletta et al., 2006). One can see that the distribution of this coefficient is similar between the central and the nearest microphone and dissimilar to that of the farthest microphone. Moreover, Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) is determined here between these distributions after fitting a five components GMM to the first order MFCC of each microphone. The divergence is found to be 0.56 between the GMMs of the central and closest microphones while it is 1.49 between the GMMs of the central and farthest microphones. Nonetheless, it might be difficult to argue that features from distant microphones can enrich statistical independence despite the evident variation in the distributions. The reason is that all microphones are simultaneously making observations about the same event.



(a) Nearest. (b) Central. (c) Farthest.

Figure 1: Distribution of first order MFCC coefficient extracted from the speech signals recorded at a central, near and a distant microphone.

Fig. 2 illustrates the correlations between the first order MFCC coefficient extracted from the signal received at the central microphone and those of the rest of the microphones individually. This is addressed for eight meeting excerpts [1] of the AMI data . Each meeting has four participants and the conversation is recorded by two circular microphone arrays. The first array has 8 microphones and is situated between the four participates. The second array has 4 microphones and it is 1.09 meters away from the first array.

Fig. 2 shows the mean correlation over these eight meetings as well as the standard deviation. It can be seen that the correlation decreases as the microphone distance from the central microphone increases. This implies that the diversity in the characteristics of the recorded speech signal seems to increase with distance. It must be noted that all coefficients of MFCC provide similar correlation pattern. The sudden change in the correlation value between the 7^{th} and the 8^{th} microphones indicates the transition in microphone selection from one microphone array to another.

This selection method does not consider the quality of the selected channels. Hence, the application of speech enhancement techniques to the microphone signals, as a pre-processing step, is necessary before speech features are extracted. In this work, a form of Wiener filtration called Two-Step Noise Reduction (TSNR) is used (Plapous et al., 2006).

2.2. Selection of Best Quality Channels

The theory behind this selection method is very different from the one behind the selection of distant microphones. This method aims to select one or more of the least distorted channels among the available ones. Reverberation can be a considerable source of distortion

[1] These meetings are identified in Section 4.1.

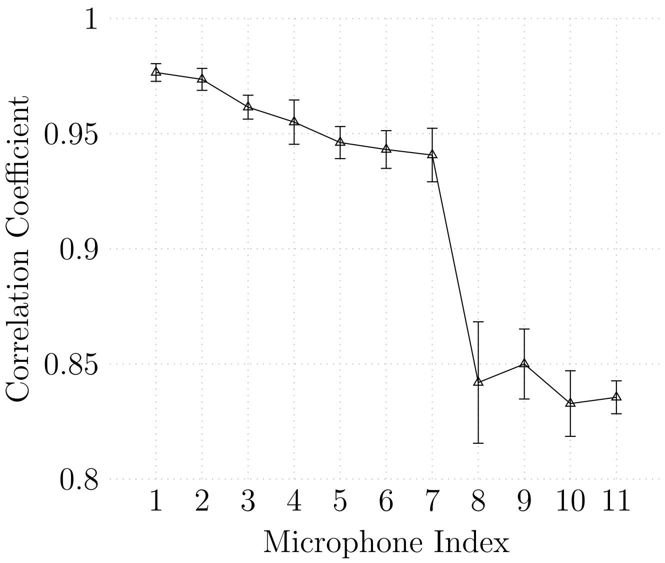


Figure 2: The correlation coefficient of channels' 1^{st} order MFCC cepstral coefficient as a function of distance from the central microphone. The correlation coefficient is determined between the 1^{st} order MFCC coefficient obtained from the central microphone and the one obtained from the first closest microphone, the second closest microphone and so on.

to the recorded speech in meeting rooms. The use of cepstral distance to identify good quality channels requires a reference channel which is assumed to provide a relatively clean speech signal. When information about the meeting room setting is provided, there might be prior knowledge about a particular microphone which can provide a good quality signal that can be used as a reference signal. It can be argued that there might be no need to conduct channel selection in such a case. In practice, however, such information is usually unavailable and the choice of a reference channel is a difficult task.

The beamformed signal obtained using the method of (Anguera et al., 2007) is proposed to be used as a reference signal here for the following reasons:

1. The segments of the signals to be combined are aligned based on estimated delays among them which strengthens the speech signal and weakens random noise effects;
2. In the final summation stage, signals' segments are weighted according to their qualities using the following equation (Anguera et al., 2007)

$$w_i(j) = \begin{cases} \dfrac{1}{M} & j = 0 \\ (1-\alpha)w_i(j-1) + \alpha R_i(j) & \text{otherwise,} \end{cases} \tag{5}$$

   where $R_i(j)$ is the average cross-correlation between segment j for channel i and the relevant aligned (based on the pre-estimated delays) segments of the rest of the channels. M is the total number of channels and α is an adaptation ratio empirically set to 0.05. One can notice that, except for $j = 0$, the calculation of the weight of any segment depends on the weight of the previous segment. However, the weight for the first segment, or when $j = 0$, is assumed to be $\frac{1}{M}$ for all of the channels.

The beamformed signal seems to be a good reference choice for the selection of the least distorted channel using cepstral distance. The beamformed signal is an enhanced signal that was found to provide better diarization performance in comparison to the signal received at the most centrally located microphone (Anguera et al., 2007). It will be shown in the results that the proposed cepstral distance based channel selection with the beamformed signal as a reference provides channel selection that has better diarization performance than the beamformed signal itself for the development data. MFCC features are used here as the cepstral representation of the speech signal in the cepstral distance calculation. The cepstral distance between two feature vectors is calculated as (Flores et al., 2018)

$$\Delta_{i,\text{ref}} = \frac{10}{\log 10}\sqrt{2\sum_{c=1}^{C} |f_i(c) - f_{\text{ref}}(c)|^2}, \tag{6}$$

where $f_i(c)$ and $f_{\text{ref}}(c)$ are MFCC cepstral coefficients of two feature vectors of channel i and the beamformed signal ref, respectively. C is the total number of MFCC cepstral coefficients (feature vector dimensionality). The term $10/\log 10$ in (6) is related to the definition of the cepstral distance as the logarithmic spectrum envelop distance (Kitawaki et al., 1982).

Let $\mathbf{X}_i$ and $\mathbf{X}_{\text{ref}}$ denote the entire set of feature vectors extracted from channel i and the beamformed signal ref. The rows of $\mathbf{X}_i$ and $\mathbf{X}_{\text{ref}}$ are the cepstral coefficients of MFCC and the columns are the feature vectors. The average cepstral distance between all feature vectors of channel i and those of the beamformed signal ref is determined as in the following

$$\tilde{\Delta}_{i,\text{ref}} = \frac{1}{T}\sum_{t=1}^{T} \frac{10}{\log 10}\sqrt{2\sum_{c=1}^{C} |\mathbf{X}_i(c,t) - \mathbf{X}_{\text{ref}}(c,t)|^2}, \tag{7}$$

where T is the number of feature vectors.

The average cepstral distance between all of the channels and the beamformed signal is calculated. Then the best channel is selected as the one that produces minimal cepstral distance from the beamformed signal ref

$$\text{best channel} = \underset{\forall i}{\arg\min}\ \ \tilde{\Delta}_{i,\text{ref}}\ , \tag{8}$$

where $i = 1, 2, ..., M$ and M is the number of microphones. Fig. 3 depicts the FFT spectrum of peer one second segments of speech from the beamformed signal as well as two channels that are selected as the best and worst channels using the proposed method. The meeting example under study is the IS1001a AMI meeting. Before making inferences about the spectrums shown, it should be noted that the regions with the highest values represent strong energy instances in speech phonemes. As such, empty (silent) regions in the spectrum should present the lowest values. These (empty) instances may not have the lowest values in

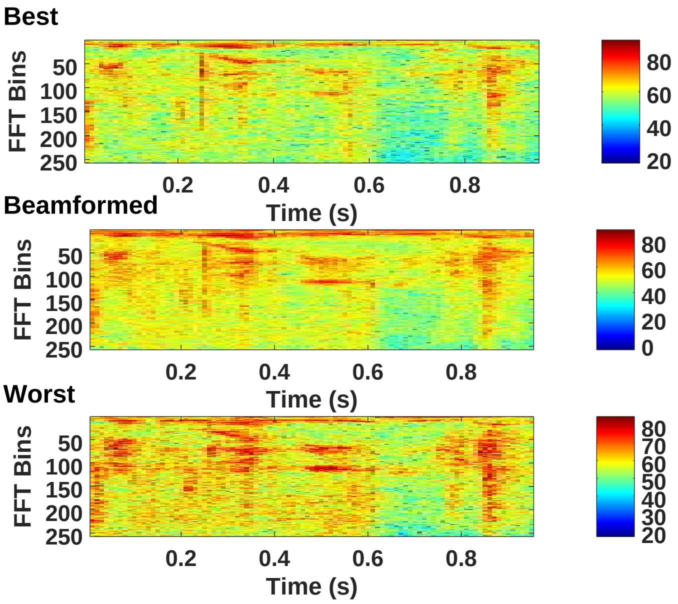


Figure 3: Spectrums of one second of speech extracted from the beamformed signal and two channels selected as the best and worst ones. This figure presents quite an informative imaging of the quality of the spectrums.

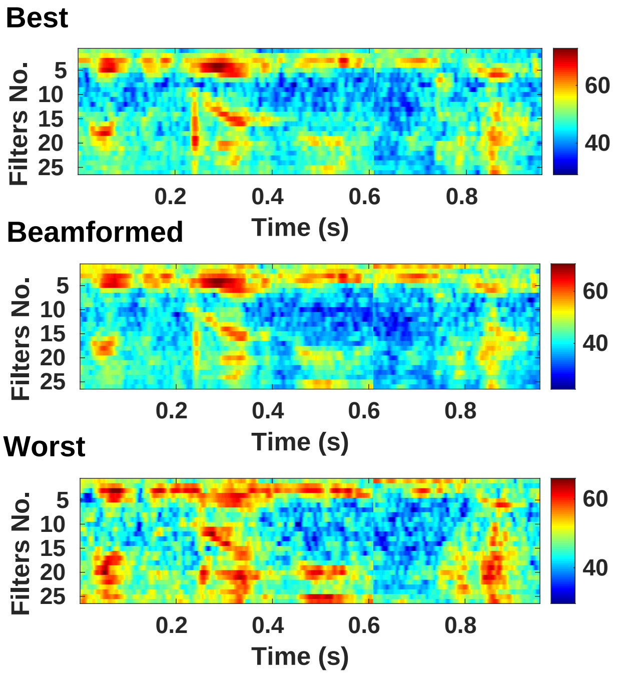


Figure 4: Insights on the filter bank decomposition of the spectrums of beast, beamformed and worst quality signals.

noisy channels, especially if the noise energy is affecting the general spectrum (for example, white noise).

It can be seen in Fig. 3 that the channel selected as the worst one exhibits what can be described as distorted speech spectrum compared to the spectrum of the beamformed signal. The dispersion of higher energy regions (spectrum smearing) suggests that this channel experiences a reverberation effect. On the other hand, the best selected channel provides a more distinct speech spectrum than the beamformed signal and the worst selected channel. It can be noticed that the beamformed signal presents higher phoneme energies than the best selected channel likely due to the alignment of segments before summation. However, it seems to have lighter coloured background than the best selected channel which indicates the presence of noise. The probable reason is that the noise is not independent between channels so that the combined effect is not fully cancelled by the weighted summation.

Additionally, Fig. 4 shows the filterbank decomposition of the spectrums selected in Fig. 3. It can be noticed that the spectrums of the best selected channel and the beamformed signal are more alike in this case. The channel selected as the worst one continued to show dispersion of high speech energy. It is evident that this selection method performs as anticipated. The beamformed signal which is used as a reference signal clearly provides a distinguishable factor between good and bad quality channels based on cepstral distance.

3. Acoustic Feature Extraction from Selected Channels' Subbands

In this method, acoustic features will be extracted from subbands of channels that are less affected by reverberation. A subband refers to a range of frequencies in the spectrum of the signal. The speech spectrum is to be first divided into a number of subbands and the reverberation effect on those subbands is characterised over all of the available channels. The least reverberated subbands among the underlying channels are then chosen. In the end, the entire speech spectrum is retained from different channels and MFCC coefficients are extracted from the log-energies of the mel-filters that correspond to each subband.

3.1. Averaged Joined Gradient Estimation

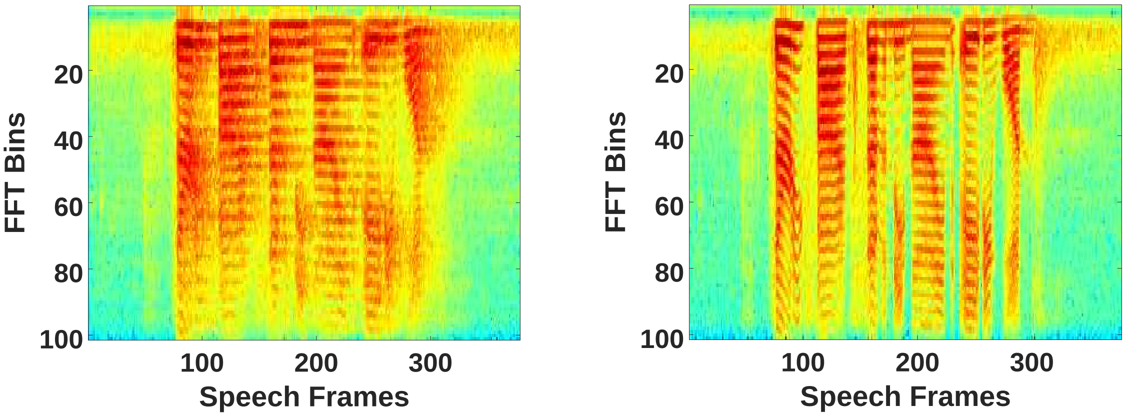
Reverberation is known to cause smearing in the speech spectrum. This effect can be visually observed when stacking together the spectral estimates of short speech frames. In the speech spectrum represented by a sequence of frames, reflected speech signal causes extensions in the speech energy from one frame to another. This is illustrated by the reverberated spectrum in Fig. 5a. On the other hand, the spectrum of clean (non reverberated) speech signal has sharper transitions from one frame to the other. The latter is illustrated in Fig. 5b.

It is proposed here, to estimate the gradients of the spectrum across the speech frames. Then this is used it to characterise the degree of reverberation. It is assumed that the smearing effect of reverberations minimises the gradient. Thus, less reverberated speech will have higher gradient values.

In order to estimate the gradient, the speech signal is first divided into frames of 25 ms length. Overlap is not allowed between frames. This is because it results in some continuity of the speech energy from one frame to another which would affect the gradient. The spectrum is determined for each frame as $\log_{10}$ of the discrete Fourier transform (FFT). In an attempt to equalise the gradient estimates over different channels; the mean and variance of the spectrum is normalised across the speech frames.

Let k denote a fraction of the spectrum (one FFT bin). Let $\eta_k(t)$ be k's value at speech frame t. For channel j, the mean of the absolute gradient of $\eta_k(t)$ for T frames is determined as

$$\xi_{k,T,j} = \frac{1}{T}\sum_{t=1}^{T} |\nabla\eta_k| \,. \tag{9}$$



(a) Reverberated Speech

(b) Clean Speech

Figure 5: Speech sample of the YOHO data (Campbell and Higgins, 1994) for a female uttering the numbers "35 79 81". Artificial reverberation of 0.7s was added to produce the reverberated sample.

The average gradient of a specific subband of the spectrum is calculated as

$$\xi_{k_1,k_2,j} = \frac{1}{k_2 - k_1 + 1}\sum_{k=k_1}^{k_2} \xi_{k,T,j}, \tag{10}$$

where k_1 and k_2 are, respectively, the low frequency edge and the high frequency edge of the subband.

Channels that exhibit high reverberation times can have similar ξ_{k_1,k_2} values which could make the degree of reverberation to be indistinguishable when measured by ξ_{k_1,k_2}. For example, it is possible that longer spread of the speech energy causes the value of ξ_{k_1,k_2} to increase which is the opposite to what was originally assumed here. The possibility of such conditions can be tackled by introducing a threshold to discard overly extended smearing of the spectrum.

The threshold is determined using all of the channels for which the reverberation to be characterised over the subband k_1 to k_2. It is calculated as

$$\iota = \frac{1}{M}\sum_{j=1}^{M} \xi_{k_1,k_2,j}, \tag{11}$$

where M is the number of channels.

This threshold is basically the average of all channel's $\xi_{k_1,k_2,j}$ values obtained in (10). The value of ι will be used to transform the gradient estimates into binary values. The new gradient estimates, or the joined gradient, will be obtained by the following transformation

$$\hat{\xi}_{k,t,j} = \begin{cases} 1 & \text{for} \quad \xi_{k,t,j} \geq \iota \\ 0 & \text{for} \quad \xi_{k,t,j} < \iota \end{cases}, \tag{12}$$

where $\xi_{k,t,j}$ is the j^{th} channel gradient value for specific fraction of the spectrum k (equivalent to an FFT bin) at frame t. Then, the new Average Joined Gradient (AJG) estimate of channel j over T frames and for a subband that extends from k_1 to k_2 is determined as

$$\hat{\xi}_{k_1,k_2,j} = \frac{1}{k_2 - k_1} \sum_{k=k_1}^{k_2} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{\xi}_{k,t,j} \right). \tag{13}$$

The higher the reverberation effect the higher the smearing it causes in the spectrum which minimises the value of $\hat{\xi}_{k_1,k_2,j}$ as assumed here. The channel that exhibits the lowest reverberation at subband k_1 to k_2 is selected using the AJG estimate of (13) as

$$j_{selected} = \arg \max_{\forall j} \hat{\xi}_{k_1,k_2,j}. \tag{14}$$

The following section introduces a framework for acoustic feature extraction based on subband selection.

3.2. Acoustic Feature Extraction Framework

As stated earlier, this selection method is designed to account for hypothetical variations in the degree of reverberation across the speech spectrum of the available channels. Three cases of subband selection will be investigated. In one case, the spectrum is divided into two equal subbands each is to be presumably selected from a different channel. The second case considers three equal subbands and the third case considers four equal subbands.

Acoustic feature extraction from selected subbands will be based on the MFCC framework. Cepstral coefficients are obtained after applying DCT to the outputs of the filters covering a particular subband. Then, cepstral coefficients of all subbands are simply concatenated. Fig. 6 illustrates the case where the spectrum is divided into three equal subbands. The filters that cover the subbands' edges will be included in the feature extraction for both of the adjacent subbands.

The maximum number of channels that can be selected is equal to the number of subbands. On the other hand, there is no restriction on having two or more subbands selected from the same channel if they are found to experience the least reverberation effect. Moreover, when a subband is selected, it is considered over the entire length of the recording.

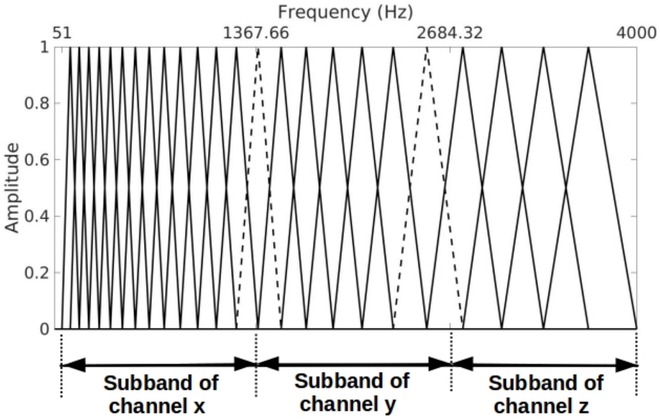


Figure 6: The framework of feature extraction from selected subbands based on MFCC methodology. Cepstral coefficients are to be calculated separately from the filters that cover each subband. Dotted filters indicate that they will be included in the feature extraction of both of the adjacent subbands.

3.3. Detection of Simulated Reverberation Effects

This subsection demonstrates the accuracy of the Average Joined Gradient (AJG) estimates in detecting the degree of reverberation. Simulated reverberation effects are added to a clean speech sample using the tool designed for the REVERB challenge by (Kinoshita et al., 2013). The tool performs convolution between the speech sample and a designated Room's Impulse Response (RIR). Three different reverberation times are added and tested:

- 0.2s using the RIR of simulation room 1 recording angle A;
- 0.5s using the RIR of simulation room 2 recording angle A;
- 0.7s using the RIR of simulation room 3 recording angle A.

Added Reverberation	$\xi_{k_1,k_2,j}$	$\hat{\xi}_{k_1,k_2,j}$
0.0 s (Original Speech)	0.356	0.603
0.2 s	0.166	0.308
0.5 s	0.159	0.276
0.7 s	0.152	0.260
Standard Deviation	0.098	0.162

Table 1: Values of the average gradient ($\xi_{k_1,k_2,j}$) and average joined gradient ($\hat{\xi}_{k_1,k_2,j}$) in relation to different added reverberation times as well as the original speech sample "21 37 63" of the YOHO data (Campbell and Higgins, 1994).

A speech sample from the YOHO data (Campbell and Higgins, 1994) for a male uttering the numbers ”21 37 63” is selected. Then its convolution with the three rooms' RIR is used to test the gradient estimations. For the entire speech spectrum, Table 1 shows the values of the average gradient ($\xi_{k_1,k_2,j}$) determined using (10) and the average joined gradient ($\hat{\xi}_{k_1,k_2,j}$) estimated using (13).

One can notice from Table 1 that both $\xi_{k_1,k_2,j}$ and $\hat{\xi}_{k_1,k_2,j}$ decrease as the reverberation time increases which accommodates the assumptions made here. The maximum values are given for non-reverberated speech. More importantly, the values of $\hat{\xi}_{k_1,k_2,j}$ have higher standard deviations which makes this measure more precise in distinguishing close reverberation times. Note that the standard deviation is determined between average gradients of different reverberation times for one sample of speech.

4. Experimental Evaluation in Binary Key based Diarization

The performance of binary key based diarization is evaluated here using the proposed methodology as its front-end. This diarization approach is fast but has somewhat non-competitive diarization accuracy which can be improved. Binary key based diarization was introduced in (Anguera and Bonastre, 2011). It is an agglomerative diarization approach where segments and clusters are modelled by fixed dimensional binary-valued vectors called binary keys, see Fig. 7. From a conversation's feature vectors, a Binary Key Background Model (KBM) is obtained as a pool of single-Gaussian models fitted to 2s segments (overlapped by 0.5s). The 896 most dissimilar Gaussians are selected using the cosine similarity among the means of the Gaussians' pool (Delgado et al., 2015a). Using the feature vectors of each segment and cluster, a binary key is obtained as follows. The log-likelihood is calculated for the feature vectors to each of the 896 single-Gaussian models of the KBM. This produces an 896 dimensional vector where each element has the log-likelihood value of the feature vectors to one Gaussian of the KBM. The elements that have values higher than a pre-defined threshold are set to 1, the rest are set to 0. This results in the binary key.

The initial clusters are obtained by flatly dividing the conversation's feature vectors into 16 equal partitions. The Jaccard coefficient is used to measure the similarity between the binary keys of each segment and each cluster. It is also used to measure the similarity between the clusters for cluster merging. For two binary keys, $\mathbf{v}_{b1}$ and $\mathbf{v}_{b2}$, the Jaccard coefficient is expressed as

$$\mathcal{J}(\mathbf{v}_{b1}, \mathbf{v}_{b2}) = \frac{\sum_{i=1}^{L} \mathbf{v}_{b1}(i) \wedge \mathbf{v}_{b2}(i)}{\sum_{i=1}^{L} \mathbf{v}_{b1}(i) \vee \mathbf{v}_{b2}(i)}, \tag{15}$$

where L is the binary key length, $\wedge$ indicates the boolean AND operator and $\vee$ indicates the boolean OR operator.

Starting from 16 initial clusters, similar segments are clustered then merging takes place. The number of clusters gradually decreases by one. The Within Cluster Sum of Squares (WCSS) is used to determine the best number of clusters (Delgado et al., 2015a). Then, a

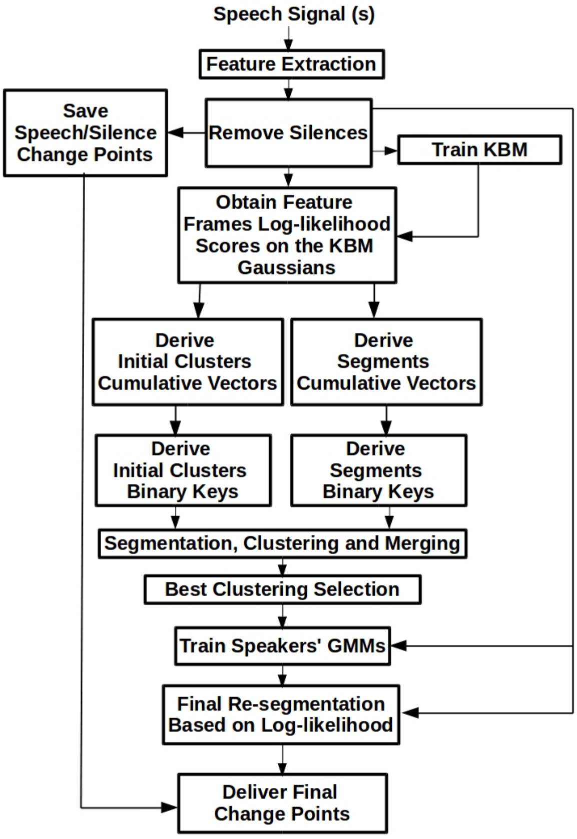

Speech Signal (s)
Feature Extraction
Save Speech/Silence Change Points
Remove Silences
Train KBM
Obtain Feature Frames Log-likelihood Scores on the KBM Gaussians
Derive Initial Clusters Cumulative Vectors
Derive Segments Cumulative Vectors
Derive Initial Clusters Binary Keys
Derive Segments Binary Keys
Segmentation, Clustering and Merging
Best Clustering Selection
Train Speakers' GMMs
Final Re-segmentation Based on Log-likelihood
Deliver Final Change Points

Figure 7: Descriptive diagram of the binary key based diarization system.

final segmentation takes place where each cluster is modelled by a 128 components Gaussian Mixture Model (GMM) fitted to its corresponding feature vectors. A 1s window is used to smooth the log-likelihood values of the conversation feature vectors to each cluster's GMM. This produces the final speakers' change points. This system configuration is fixed and only the input features change as it will be shown shortly. The performance is evaluated in terms of Diarization Error Rate (DER) and Speaker Error Rate (SER) as defined in (Anguera, 2006). Note that $\mathrm{DER_b}$ and $\mathrm{SER_b}$ in the following tables represent baseline DER and SER, respectively.

4.1. Baseline System Performance

This subsection reports the baseline system performance for the corpora under investigation. The corpora consists of 16 meeting excerpts of the AMI corpus (Carletta et al., 2006). The first eight meetings, referred to as the IS1000 set, include meetings: IS1001a, IS1002d, IS1003a, IS1004a, IS1005a, IS1006a, IS1007a and IS1009a. Each meeting includes four speakers and is recorded by 12 microphones. The second eight meetings, referred to as the TS3000 set, include meetings: TS3004a, TS3005a, TS3006a, TS3007a, TS3008a, TS3009a, TS3010a and TS3011a. Each meeting includes four speakers and is recorded by 18 microphones. The final evaluation will be carried on the RT-05S set (Fiscus et al., 2005) and the evaluation subset of the CHiME-6 Track 2 dataset (Watanabe et al., 2020). The RT-05S set consists of ten meetings recorded by different number of microphones and include variable number of speakers. The CHiME-6 evaluation set consists of two dinner parties each is recorded by 24 microphones and includes four participants. In all cases, the number of speakers is estimated by the system and not provided as an input.

The system uses conventional MFCC features (Hamming window) extracted from the beamformed signal of each meeting excerpt. All available channel signals are used in the beamforming process. The first best delays selected by the Viterbi algorithm are used in the alignment of the segments. Delays and channel weights are estimated every 250 ms for 500 ms segment size. The MFCC features are extracted from the beamformed signal for speech frames of 25 ms in size at 10 ms rate (every 10 ms). The number of filters in the filter bank are 24 and the number of cepstral coefficients are 19 excluding the 0^{th} order coefficient. Non-speech feature vectors are identified using the reference files associated with each excerpt in order to precisely evaluate the diarization performance as in (Delgado et al., 2015a). The baseline performance is reported in Table 2.

Dataset	**$\mathrm{DER_b}$** (%)	**$\mathrm{SER_b}$** (%)
IS1000 Set	36.41	35.90
TS3000 Set	41.25	40.30
RT-05S Set	30.90	21.30
CHiME-6 Eval Set	63.84	49.20

Table 2: Baseline System Performance. By the end of this section, considerable relative improvements will be shown compared to this baseline performance that only uses MFCC features extracted from beamformed signals.

4.2. System Performance for Acoustic Features Extracted from Selected Channels

This subsection presents a study of system performance in light of the proposed channel selection methods. The extraction parameters of MFCC, as acoustic features, are the same as those of the baseline system. Initially, for the IS1000 meetings used here as a development set, Table 3 shows the effect of using a concatenation of features extracted from each individual channel in comparison to those extracted from the beamformed signal.

Signal	**Feature Dim.**	**DER (%)**	**SER (%)**
Beamformed (baseline)	19	36.41	35.90
All Channels	228	28.06	27.60

Table 3: Performance comparison between the case of MFCC features extracted from the beamformed signal and a concatenation of MFCC features extracted from each channel for the IS1000 development set.

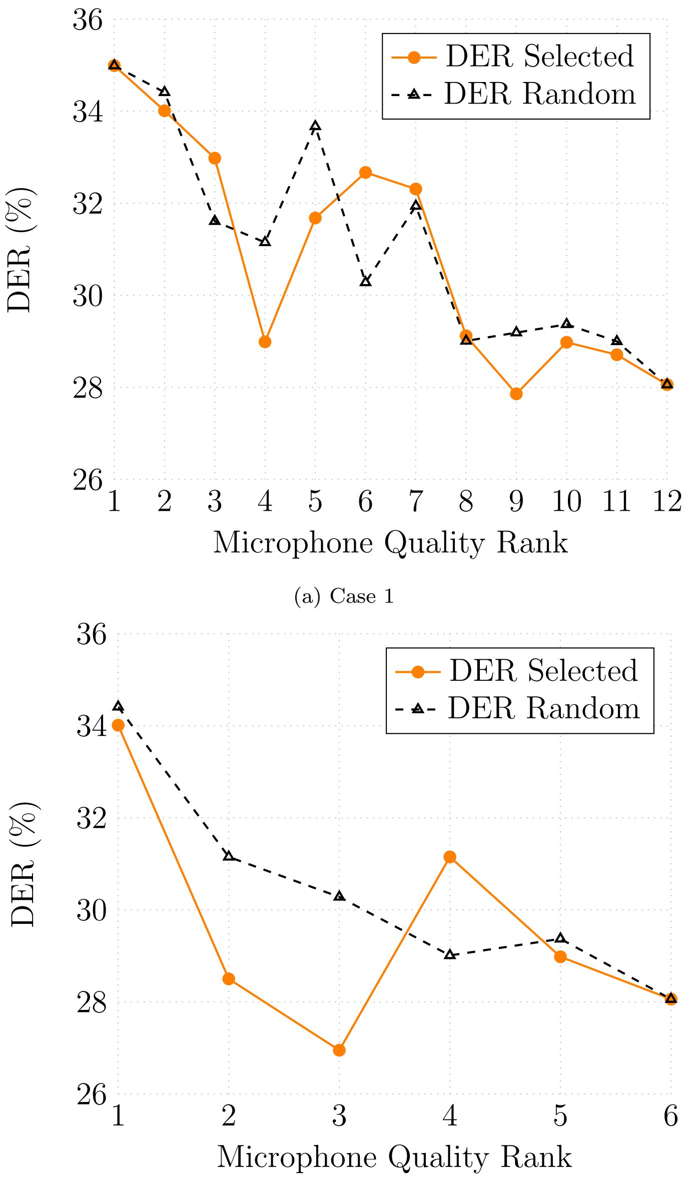
A considerable reduction in DER and SER can be noticed in Table 3 as an effect of using a concatenation of all microphones MFCC features. However, as mentioned earlier, concatenation of all features comes at the cost of increasing the dimensionality. For the IS1000 development set, where each excerpt is recorded using 12 microphones, MFCC feature dimensionality grew from 19 (of the single beamformed signal) to 228 which increases the processing time. Accordingly, the channel selection methods proposed attempts to achieve similar improvement in the performance but with a lower computational load.

4.2.1. Distant Channels

The IS1000 development set is used to approximate the suitable number of selected channels that provide an improvement in the diarization accuracy. This channel selection method cannot be evaluated on the RT-05S NIST evaluation set because there are four meetings with three channels and less. However, the final outcome of the method is tested on the CHiME-6 evaluation set in addition to the TS3000 set.

The diarization performance is investigated for two cases of this selection method. The first case is the selection of the most centrally located microphone and then additional channels are selected starting from the most distant one from the central microphone. The second case is where two groups of distant channels are selected. One group includes the central microphone and the ones close to it and the second group is one that is distant from the first group. In this case, the selection also starts from concatenating the features from the centrally located microphone and the most distant one. Then, features of two more channels are added, one of them is the nearest to the central one and the other is the second most distant one, and so on.

Fig. 8a demonstrates the system performance in terms of DER in relation to the first case of this selection method. A trade off must be made between dimensionality growth and system performance. It can be noticed that concatenating features of three distant channels in addition to the central one (total of 4) provided equivalent performance to the one when all channels' features are used (Table 3). After that, an increase in the errors can be noticed supposedly due to decrease in diversity. Then, as expected, the performance moves toward



(a) Case 1

(b) Case 2

Figure 8: Effect of distant microphones selection (AMI IS1000 set) on DER. The figures contrast the performance of the selection criteria (DER Selected) with the case of using randomly selected microphones in addition to the centrally located one (indicated by DER Random). Case 1: microphone index 1 means that only features of the central microphone are used. Then, features of the rest of the channels are added starting from the most distant channel. Thus, the horizontal axis indicates the total number of microphones. Case 2: features of two groups of distant channels are used. Index 1 means the pair of the central and the most distant channels.

the one achieved when all channels' features are used. One might notice that the lowest error occurred at nine selected channels, however, it is not a favourable operation point given the high dimensionality of features at that case. One may also notice from Fig. 8a that using the most centrally located microphone alone provided better performance than the one when the beamformed signal is used, refer to Table 2. This can be the case sometimes since the most centrally located channel is also considered a good channel for feature extraction, see (Anguera et al., 2007).

Fig. 8b demonstrates the second case of selecting two groups of distant microphones. It can be observed that the choice of three pairs (distant groups of three microphones each) provides an appealing trade off between the performance and the number of channels (six in total). The DER at this point is also lower than the case of concatenating all channels' features. The case of Fig. 8b, better demonstrates the achievement of the desirable diversity using this selection criterion. Selecting two distant groups of microphones is assumed to provide more diversity between the channels, compared to the case of Fig. 8a, hence, better performance is achieved.

It is evident, from Fig. 8b, that using a concatenation of features of only one pair of distant microphones provides a marginal 2% decrease in DER compared to the case of using only features extracted from the beamformed signal (Table 2). This can also confirm the basis of this selection method. For the AMI evaluation data (TS3000) there was a decrease of about 4% in DER using features extracted from a central and one distant microphones.

The theoretical behaviour of the plots of Fig. 8 was anticipated to be as in the following. A concatenation of a few number of distant channels improves the performance as the diversity is assumed to be high. By adding more channels, the errors are expected to increase as a result of decrease in the diversity. Then the performance is supposed to improve again by adding more channels as it approaches the case of concatenating all channels' features. The plots of Fig. 8 fairly accommodated the expectations. Sharp changes occurred due to the fact that the DER and SER depend on the outcomes of three components within the system: clustering, best clustering selection and the final re-segmentation. In the process of changing the amount of features, a small variation in one component's outcome can cause non-smooth changes in the subsequent ones.

Tables 4 and 5 show system performance in light of both cases of the proposed channel selection method at the points that gave the lowest errors based on the IS1000 development set. By comparing the results in tables 4 and 5 to those of Table 2, one can notice that there is a maximum relative improvement in DER of around 8% on the TS3000 evaluation set. While the development set exhibited a maximum relative improvement in DER of about 25%. It is normal that the relative improvements on the IS1000 development set and TS3000 evaluation set are different because of the difference in meetings conditions. However, this selection method provides a cost effective alternative to beamforming. Using features of only one pair of distant microphones, this method presented relative improvements of 6.59% and 10.15% for the IS1000 development set and TS3000 evaluation set, respectively.

Dataset	**DER (%)**	**SER (%)**	**DER$_\mathbf{b}$ (%)**	**SER$_\mathbf{b}$ (%)**
IS1000 Dev Set	28.99	28.50	36.41	35.90
TS3000 Eval Set	38.03	37.10	41.25	40.30

Table 4: System performance for the IS1000 development set and TS3000 evaluation set as an effect of features concatenation of one central channel and three distant channels.

Dataset	**DER (%)**	**SER (%)**	**DER$_\mathbf{b}$ (%)**	**SER$_\mathbf{b}$ (%)**
IS1000 Dev Set	26.95	26.50	36.41	35.90
TS3000 Eval Set	38.87	37.90	41.25	40.30
CHiME-6 Eval Set	61.75	47.10	63.84	49.20

Table 5: System performance for the IS1000 development set and the TS3000 and CHiME-6 evaluation sets as an effect of features concatenation of two distant groups of channels. Each group has three microphones.

4.2.2. Best Quality Channels

The beamformed signal is used here as a reference to assess the quality of the channels using the cepstral distance and based on 19 MFCC coefficients extracted from the available speech signals.

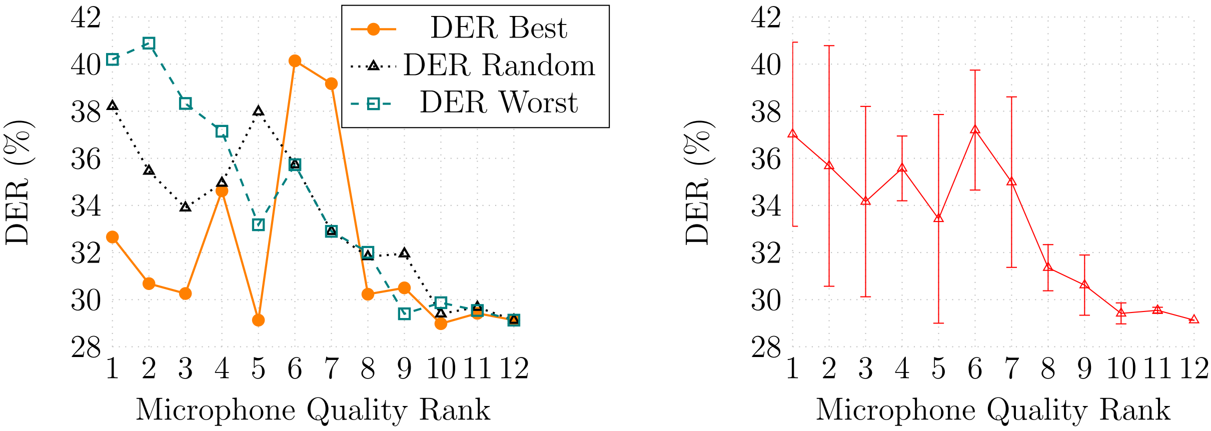
The performance of speaker diarization based on acoustic features from channels selected with this method is investigated using the IS1000 development data. The development data is used to estimate the sufficient number of best quality channels which the concatenation of their features improves the performance. In this method, the channels' signals are not processed by any form of speech enhancement techniques. Therefore, the quality of the channels is assessed using the proposed method without any speech enhancement.

Fig. 9a shows the changes in DER in relation to varying the number of best selected channels. Concatenated features of these best selected channels are used in binary key based diarization. In the figure, the diarization performance using features of the first selected best channel is superior to that of the beamformed signal. This channel presented lower DER by about 4%. Then concatenating additional features of more good quality channels decreased the DER even further. The increase in DER at 4 channels is caused by a sudden increase in the DER of meeting IS1001a. For this meeting, the DER was 39.94% for three channels, then it jumped to 61.34% for four channels and then it returned to 39.00% with five channels. When the DER was 61.34%, the system mistakenly estimated that the number of speakers in meeting IS1001a is eight, while the actual number of speakers is only four.

The lowest DER is achieved with five channels. This best performance point is followed by an increase in DER, possibly as a result of adding features from poor quality channels. Then the error decreases again, potentially, because of the increased feature dimensionality and diversity. The error continues to decrease as a result of approaching the performance when all channels' features are used, see Table 3. For comparison, Fig. 9a also presents results on the concatenation of features of random channels in addition to a selection that starts with worst quality channels. Worst quality channels are selected as the ones that have the highest cepstral distance from the beamformed signal. One can notice from the figure that the curve for worst quality channels provides degraded performance in the beginning. Then, it improves when the number of channels increases. On the other hand, the curve of the random selection has no particular trend. In all cases, the performance improves when the number of channels increases. However, the relatively high increase in dimensionality causes the processing time to increase.

Fig. 9b shows the standard deviation (SD) of the values of the curves in Fig. 9a. At the beginning, high SD can be noticed due to the relatively high differences in the values of best and worst quality channel selection. The situation in the case of four channels happened because of the reasons explained above regarding the IS1001a meeting. SD then decreases in the case of six channels and it was expected to continue decreasing until it reaches zero at the case of twelve channels, which is almost the case. However, the case of seven channels has higher SD than the case of six channels; nonetheless, it has less SD than the cases of one, two, three and five channels. The relatively high SD in the case of seven channels is partially caused by improvements in the cases of worst and random channel selection.

By using a concatenation of features in the diarization, it can be difficult to tell if this selection method is performing as anticipated. It can be more informative to report the system performance, for illustrative purposes, when features of the selected channels are individually used in the diarization. Fig. 10 demonstrates the system performance in relation to using features from individual channels, starting from the best quality channel as selected by this method. The curve of Fig. 10 implies that the channel quality estimation process is performing fairly as anticipated. In general, the DER increases as the quality of



(a) Different quality selection setups.

(b) Mean and standard deviation of Fig. 9a.

Figure 9: Effect of best quality microphones selection (IS1000 development set) on DER. In Fig. 9a, DER-Best indicates the case when the concatenation start from the best channels. DER-Worst indicates the opposite case. DER-Random indicates random selection. Fig. 9b represents the mean and standard deviation of the values of DER-Best, DER-Worst and DER-Random.

the channel decreases. However, the concatenation of features from good quality channels results in lower DER as reported in Fig. 9a.

Given the similar level of errors for the range of channels 7 to 11 (Fig. 10), it can be inferred that the qualities of those channels are similar. Therefore, those low quality channels were not optimally ranked by the selection method minimising the impact on the final results. Thus, it appears that the method provides good selection performance, in general, in addition to identifying the worst quality channel (channel 12) which provides the highest error.

In the evaluation, features concatenated from up to five of the best channels are used with the TS3000, the RT-05S NIST and the CHiME-6 evaluation sets. Table 6 shows the results for these three sets in addition to the development set. The proposed method of best quality channel selection presents a relative improvement in DER of about 20% for the IS1000 development set and 8% for the CHiME-6 evaluation set in comparison to the case of using MFCC features extracted from the beamformed signals. For the RT-05S NIST evaluation set, the proposed method introduces a relative improvement of 14.43% in DER and 20.65% in SER.

Dataset	**DER (%)**	**SER (%)**	**DER$_b$ (%)**	**SER$_b$ (%)**
IS1000 Dev Set	29.13	28.60	36.41	35.90
TS3000 Eval Set	38.84	37.90	41.25	40.30
RT-05S NIST Eval Set	26.44	16.90	30.90	21.30
CHiME-6 Eval Set	58.58	43.90	63.84	49.20

Table 6: System performance for the IS1000 development set and for the TS3000, RT-05S NIST and CHiME-6 evaluation sets as an effect of features concatenation of a maximum of five best quality channels.

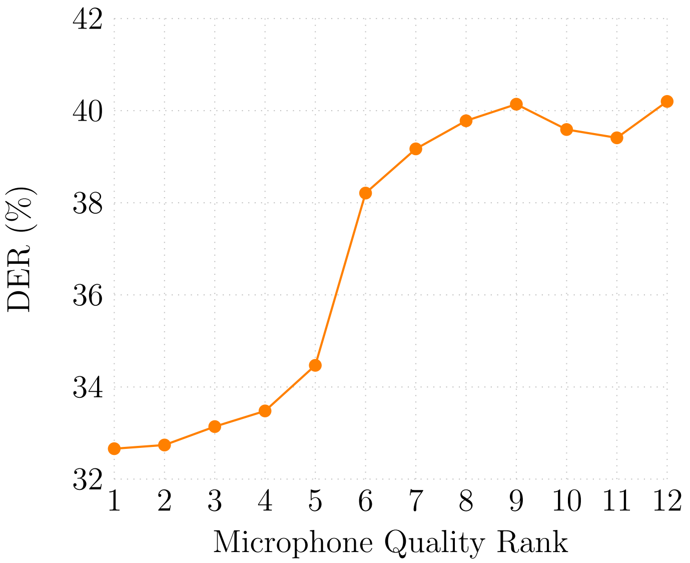


Figure 10: Effect of microphones selection (IS1000 development set) on DER starting from the best quality microphone. This plot demonstrates the efficiency of this selection method in distinguishing signals' qualities, specifically, for the range between 1 and 9.

4.3. Performance for Acoustic Features Extracted from Selected Channels Subbands

The three cases of subbands division to be investigated are summarised in Table 7. Despite that the proposed subband selection method aims to account for possible differences in reverberation effects over the subbands, the extent of reverberation over the spectrum is unknown. In an attempt to tackle such uncertainty when examining a subband, the AJG ($\hat{\xi}$) will be estimated over the subband plus 50% extensions with the adjacent subbands as described in the second column of Table 7.

The choice of the best subband division case will be made based on experiments performed using the 16 meeting excerpts of the AMI corpus. In other words, both of the IS1000 and TS3000 sets will be used as development data. New baseline performance is produced here using features extracted from beamformed signals. The reasons is to make the comparison consistent with the feature extraction setup of this method. In this method, the number of cepstral coefficients that are kept after applying Discrete Cosine Transform (DCT) is the number of filters minus 1, where only the 0^{th} coefficients are discarded. Thus, 23 MFCC coefficients are obtained after extraction with 24 mel-filters filter bank. The new baseline results using these 23 MFCC coefficients, DER_{nb} and SER_{nb}, are shown in Table 8.

Table 9 shows the system performance using the combination of IS1000 & TS3000 sets for the 2, 3 and 4 division cases of the subband selection and feature extraction process describe here (as summarised in Table 7). Both of the 2 and 3 subband cases improve the accuracy over the reference performance shown in Table 8. In theory, having smaller sections of the spectrum selected from different channels is not expected to degrade the performance. However, the case of 4 subbands has slightly degraded the performance which can be a limitation of the feature extraction framework. In the case of 4 subbands, the

No. of Subbands	Spectrum Limits for Estimating $\hat{\xi}_{k_1,k_2,j}$	Subbands for Feature Extraction	Mel-Filters Subset
2	51 - 3012.5 Hz,	51 - 2025 Hz,	1 - 17,
	1037.5 - 4000 Hz.	2025 - 4000 Hz.	17 - 24.
3	51 - 2025 Hz,	51 - 1367.66 Hz,	1 - 14,
	709.33 - 3342.65 Hz,	1367.66 - 2684.32 Hz,	14 - 20,
	2025 - 4000 Hz.	2684.32 - 4000 Hz.	20 - 24.
4	51 - 1531.25 Hz,	51 - 1037.5 Hz,	1 - 11,
	543.75 - 2518.75 Hz,	1037.5 - 2025 Hz,	11 - 17,
	1531.25 - 3506.25 Hz,	2025 - 3012.5 Hz,	17 - 21,
	2518.75 - 4000 Hz.	3012.5 - 4000 Hz.	21 - 24.

Table 7: Summary of the MFCC based feature extraction framework from selected channels' subbands. The third column shows the subbands to be selected from different channels. The exact subband of a channel used in the feature extraction can be slightly extended as a result of the actual number of filters used, refer to Fig. 6. The feature dimensionality is 23 for all cases.

Dataset	$\mathbf{DER_{nb}}$ **(%)**	$\mathbf{SER_{nb}}$ **(%)**
IS1000	34.66	34.20
TS3000	44.68	43.70
IS1000 & TS3000	39.99	39.20
RT-05S	32.13	22.60
CHiME-6 Eval	64.58	49.90

Table 8: Binary key based system performance for the datasets under investigation using 23 dimensional MFCC features extracted from beamformed signals. These results are the baseline to which the subband selection results are compared.

third subband (2025 - 3012.5 Hz) is decomposed using 5 mel-filters and the fourth subband (3012.5 - 4000 Hz) is decomposed using 4 mel-filters. Since those are overlapped filters, they can have poor transformation of the spectrum because they are expected to have relatively high residual in the correlation matrix of their log-energies as discussed in (Ahmed et al., 2019).

From Table 9, one can notice that the best results are obtained for the case of three subbands. This finding is further investigated and evaluated on the RT-05S and CHiME-6 evaluation sets and the results are shown in Table 10. The same table also presents separate results for each of the IS1000 and the TS3000 sets. The results of Table 10 appear to show a noticeable relative improvement over the case of using MFCC features extracted

No. of Subbands	DER (%)	SER (%)
2	38.87	38.10
3	35.02	34.30
4	40.61	39.90

Table 9: Binary key based system performance for the combination of IS1000 and TS3000 sets for the cases of 2, 3 and 4 subbands.

from the beamformed signal (Table 8). The results appear to provide evidence that the methodology presented in this section might considered to be a better practice than the process of combining all channels' signals into a single beamformed signal. Beamforming is a time domain process that does not take into account the spectral properties for individual microphones and in particular any deficiencies in specific ranges of a microphone's signal spectrum. These results appear to show that channels' subbands selection discards channels with spectrums that may have been distorted by reverberation effects or other degradation.

Dataset	DER (%)	SER (%)	$\mathbf{DER_{nb}}$ (%)	$\mathbf{SER_{nb}}$ (%)
IS1000	30.00	29.50	34.66	34.20
TS3000	39.45	38.50	44.68	43.70
RT-05S	28.21	18.60	32.13	22.60
CHiME-6 Eval	56.37	41.70	64.58	49.90

Table 10: The performance of the binary key based diarization system for each of the IS1000, TS3000, RT-05S and CHiME-6 sets in the case of three equal subbands spectrum division.

5. Experimental Evaluation in GMM-BIC based Diarization

The purpose of the evaluations provided in this section is to robustly validate the methods proposed in this work. This is achieved by assessing the performance of a different speaker diarization framework using the front-ends presented here. The well-known diarization approach introduced by (Ajmera and Wooters, 2003) is deployed in this framework. This diarization system is commonly used to obtain state-of-the-art performance as in (Martínez-González et al., 2017) and as a reference as in (Delgado et al., 2015b). It fundamentally depends on the Bayesian Information Criterion (BIC) to make cluster merging and merging stopping decisions, see (Ajmera and Wooters, 2003).

This system has been widely used and extensively described in the literature, see (Pardo et al., 2007b). The underlying conception models an utterance (segment of the conversation) using an ergodic Hidden Markov Model (HMM) whose number of states is equal to the number of initial clusters. The substates in this HMM impose a fixed duration on the cluster and they share a Probability Density Function (PDF) modelled by a Gaussian Mixture model (GMM) with a diagonal covariance matrix. Practically, the system performs feature vector based GMM segmentation and BIC clustering. The operation of this system as used here is described as follows. Feature vectors corresponding to silences are first removed. The rest of the feature vectors are uniformly divided into K sections, each one represents an initial cluster. Each cluster is modelled by a five components GMM.

The conversation feature vectors are also segmented into portions of 2.5 seconds which is the fixed duration imposed here. Then, a number of GMM models training and segments assignment are carried which involves the Viterbi decoding followed by a merging step. When cluster merging takes place, the number of GMM components of the new cluster is the sum of the number of components of the original clusters' GMMs. Two clusters are merged if they have the greatest ΔBIC, where $\Delta\text{BIC} = \log p(D|\theta) - \log p(D_a|\theta_a) - \log(D_b|\theta_b)$. θ_a and θ_b are the GMMs of clusters a and b, respectively. D_a and D_b are the feature vectors that are used to fit θ_a and θ_b, respectively. D is the union of D_a and D_b and it is used to fit the model θ of the prospective newly merged cluster. Cluster merging stops when ΔBIC for all remaining cluster pairs is less than zero. In that case, a final re-segmentation is carried using Viterbi decoding where the segment size is reduced to 1.25 seconds. This is to allow the detection of shorter change points between speakers. In the Viterbi decoding, the transition matrix is always set manually, where $\alpha = \beta = 1$ as in (Anguera, 2006). Thus, the values of the elements of the transition matrix diagonal are equal to α, whereas the values of the off-diagonal elements are equal to $\frac{\beta}{K-1}$.

The pre-described configuration is used here to produce baseline performance. The IS1000 dataset is used as a development set to optimise these parameters. The baseline results, see Table 11, for the four datasets (IS1000, TS3000, RT-05S and CHiME-6) are achieved using features (19 MFCCs) extracted from the beamformed signals. The front-ends proposed are used as previously optimised in the binary key based approach. This means the features of a maximum of five best channel selection, of two groups of three distant microphones and of selected three least reverberated subbands.

However, a number of parameter optimisation are carried on the diarization system of this section to achieve the best results. For this purpose, the IS1000 dataset is used as a development set. In the new configuration, the number of initial clusters is $K = 20$. The segment duration is 1 second in both of the iterative and final re-segmentation processes. Moreover, the concatenation of multiple channels' features, in the case of best and distant channel selection, is found to harm the system performance. This is due to the increase in feature dimensionality (up to 114) which resulted in poor cluster modelling using the GMMs. This is because of the relatively small amount of data used to fit the GMMs.

Two different solutions are adopted to overcome this problem. In one case, Principal Component Analysis (PCA) is used to reduce the dimensionality of the concatenated features to only 19 coefficients as in the baseline case. In the other case, multiple sub-systems are used where each sub-system operate on the features of one channel. Equal-weight score fusion is used to relate the diarization process and final output. Table 11 reports all the results of this section. Noticeable relative improvements can be seen in the table on the datasets used in the experiments. Table 12 presents a comparison to the results reported in the literature on the standard RT-05S dataset. The methods proposed in this work enabled the GMM-BIC based system to outperform a number of diarization approaches from the

System Configuration	**IS1000 (Dev. Set)**		**RT-05S**		**TS3000**		**CHiME-6 Eval**	
& Front-End Type	**DER**	**SER**	**DER**	**SER**	**DER**	**SER**	**DER**	**SER**
Baseline	30.80	30.30	27.44	17.90	38.28	37.30	59.75	45.10
Best Channels Selection								
Multiple Sub-Systems	22.43	21.90	24.18	14.60	31.78	30.80	53.27	38.60
Single System	**20.60**	**20.10**	**23.95**	**14.40**	**30.35**	**29.40**	55.39	40.70
Distant Channels Selection								
Multiple Sub-Systems	28.87	28.40	–	–	37.20	36.20	53.43	38.70
Single System	22.85	22.40	–	–	33.79	32.80	**52.83**	**38.10**
Channels Subband Selection	23.94	23.40	25.00	15.40	33.19	32.20	54.04	39.40

Table 11: This table shows the performance of the GMM-BIC based diarization system subject to the front-ends proposed in this work. The baseline performance of the standard MFCC coefficients that are extracted from beamformed signals is also reported for comparison.

literature (see Table 12).

System	**Front-End**	**DER (%)**	**SER (%)**
GMM-BIC based Multiple Sub-Systems	Acoustic (Best channel selection)	**23.95**	**14.40**
Information Bottleneck (Vijayasenan et al., 2008)	Acoustic + Spatial	–	17.70
GMM-BIC based System (Anguera and Bonastre, 2011)	Acoustic	24.96	–
Online i-vector with Information Bottleneck (Madikeri et al., 2015)	Acoustic	–	16.10
PLDA i-vector with Information Bottleneck (Madikeri et al., 2015)	Acoustic	–	16.50
Robust GMM based Modelling (Peso, 2016)	Spatial	–	17.00

Table 12: Summary of the performance of a number of diarization systems for the RT-05S NIST set compared to the best result achieved in this work.

Moreover, the best result presented in our work using the CHiME-6 evaluation dataset is achieved within the GMM-BIC based framework using MFCC features of six distant channels as a front-end. The recent work in (Medennikov et al., 2020) carried evaluations of the performance of a number of diarization approaches using the CHiME-6 Track 2 evaluation set. Within a number of frameworks, the work considered various front-ends such as MFCC, i-vectors, x-vectors and Wide ResNet (WRN) x-vectors. The diarization frameworks included Agglomerative Hierarchical Clustering (AHC) based on Probabilistic Linear Discriminant Analysis (PLDA) scoring as well as Spectral Clustering (SC) based on cosine similarity scoring. They also included End-to-End Neural Diarization (EEND) and a novel Target-Speaker Voice Activity Detection (TS-VAD) approach that is based on a single channel (TS-VAD-1C) and multiple channels (TS-VAD-MC).

Table 13 provides a comparison of this result to those presented in (Medennikov et al., 2020). It can be noticed that the DER in our framework is lower than most of those reported in the table except for the DER of the TS-VAD approach. This is probably because TS-VAD uses two different front-ends and that its configuration is based on the actual number of speakers. More importantly, the TS-VAD approach is designed to address the problem of overlapped speech which is a major issue in the CHiME-6 dataset.

System	**Front-End**	**DER (%)**
PLDA based AHC	x-vectors	68.20
PLDA based AHC	WRN x-vectors	63.79
Cosine Similarity based SC	WRN x-vectors	60.10
EEND	WRN x-vectors	56.01
GMM-BIC based AHC	**Distant Ch. MFCC**	52.83
TS-VAD-1C	MFCC + i-vectors	39.80
TS-VAD-MC	MFCC + i-vectors	37.57

Table 13: Summary of the performance of a number of diarization systems (see (Medennikov et al., 2020)) for the CHiME-6 Track 2 evaluation set compared to the best result achieved in this work.

6. Conclusions

The binary key diarization framework is relatively fast. It can complete the diarization process of an entire recording within 3.5% of its duration (Delgado et al., 2015b). Thus, it is preferred here for the purpose of optimising the performance of the proposed methods. Additionally, when feature dimensionality increases due to concatenation of channels, binary key based diarization performance can still be considered fast. This is because acoustic features are projected once on the KBM. Afterwards, the rest of the process is based on the obtained binary keys which are compared against each other using the Jaccard coefficient.

The selection of distant microphones appears to be functional in achieving diversity among the chosen channels. Variations in microphones locations can affect the underlying physical phenomena governing the propagation of sound in a room. For example, the interaction of the speech wave-front with obstacles varies depending on the path of propagation. Therefore, distant channel selection should introduce diversity in the conditions of the speech recorded by microphones of different locations. It presents a best relative improvement in the performance of the GMM-BIC based system of 25% and 11% for the development and evaluation sets, respectively.

Channel selection is normally used to exclude degraded channels in speech processing. The cepstral distance approximates the distance between the log-spectra of two speech frames represented by the cepstral coefficients. Therefore, with the beamformed signal as a reference, the cepstral distance provides the anticipated outcome of good quality channel selection. It is found to successfully identify good quality channels. This selection provides a maximum relative improvement in the performance of the GMM-BIC based system of 33% and 20% for the development and evaluation sets, respectively.

Furthermore, the subband selection introduced here targets degradation exhibited by a channel across a specific range of its spectrum. Subband selection particularly addresses the typical problematic effect of reverberation. The average joined gradient measure presented is found to be successful in differentiating the amount of reverberation that affects the speech spectrum. It is feasible in different recording conditions because it depends on a threshold that is calculated from the signals to be compared. In comparison to good quality selection, subband selection can ensure the exclusion of degraded subbands. Those subbands can be more important for speaker discrimination and are otherwise selected from a different channel. Subband selection presented a maximum relative improvement in the performance of the GMM-BIC system of 22% and 13% for the development and and evaluation sets, respectively.

The experimental results introduced in this paper shows that the presented approaches outperform acoustic beamforming. This implies that a more efficient exploitation of microphone signals is achieved here. Despite that subband selection can be more thorough than good quality channel selection, it can be noticed that the latter presents better performance. This can be attributed to the difference in feature dimensionality. However, when feature dimensionality is an issue in some scenarios, subband selection can provide an efficient solution and satisfying improvement.

A future work may allow different channels to be used for acoustic feature extraction at different times. Such paradigm would address the potential effect of speaker movements on channel selection and, consequently, on the extracted features and system performance. The effect of speaker movements can be more pronounced when channel subband selection is engaged. Moreover, the system performance can be examined when selected channels are combined using, for example, beamforming before feature extraction. This can be useful in the case of GMM-BIC based system as an alternative to using multiple systems or to using PCA on concatenated features.

Acknowledgements

The first author would like to thank the Higher Committee for Education Development in Iraq for funding his PhD study at the University of Portsmouth.

References

Ahmed, A. I., Chiverton, J., Ndzi, D., Becerra, V., 2019. Speaker recognition using pca-based feature transformation. Speech Communication.
URL http://www.sciencedirect.com/science/article/pii/S0167639318301031

Ajmera, J., Wooters, C., November 2003. A robust speaker clustering algorithm. In: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721). St Thomas, VI, USA, USA, pp. 411–416.

Anguera, X., 2006. Robust speaker diarization for meetings. Universitat Politècnica de Catalunya.

Anguera, X., Bonastre, J.-F., 2011. Fast speaker diarization based on binary keys. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, pp. 4428–4431.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., 2012. Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing 20 (2), 356–370.

Anguera, X., Wooters, C., Hernando, J., 2007. Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech, and Language Processing 15 (7), 2011–2022.

Bosworth, B. T., Bernecky, W. R., Nickila, J. D., Adal, B., Carter, G. C., October 2008. Estimating signal-to-noise ratio (SNR). IEEE Journal of Oceanic Engineering 33 (4), 414–418.

Campbell, J., Higgins, A., 1994. YOHO speaker verification corpus LDC94S16. Available at the LDC website: http://www. ldc. upenn. edu.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P., 2006. The AMI meeting corpus: A pre-announcement. In: Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction. MLMI'05. Springer-Verlag, Edinburgh, UK, pp. 28–39.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing 28 (4), 357–366.

Dawalatabad, N., Madikeri, S. R., Sekhar, C. C., Murthy, H. A., April 2016. Two-pass ib based speaker diarization system using meeting-specific ann based features. In: INTERSPEECH. San Francisco, USA, pp. 2199–2203.

Deco, G., Obradovic, D., 2012. An information-theoretic approach to neural computing. Springer Science & Business Media.

Delgado, H., Anguera, X., Fredouille, C., Serrano, J., 2015a. Fast single-and cross-show speaker diarization using binary key speaker modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (12), 2286–2297.

Delgado, H., Anguera, X., Fredouille, C., Serrano, J., August 2015b. Improved binary key speaker diarization system. In: 2015 23rd European Signal Processing Conference (EUSIPCO). Nice, France, pp. 2087–2091.

Falk, T. H., Zheng, C., Chan, W., Sept 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Transactions on Audio, Speech, and Language Processing 18 (7), 1766–1774.

Feng, X., Zhang, Y., Glass, J., May 2014. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, pp. 1759–1763.

Fiscus, J. G., Radde, N., Garofolo, J. S., Le, A., Ajot, J., Laprun, C., 2005. The rich transcription 2005 spring meeting recognition evaluation. In: International Workshop on Machine Learning for Multimodal Interaction. Springer, pp. 369–389.

Flores, C. G., Tryfou, G., Omologo, M., 2018. Cepstral distance based channel selection for distant speech recognition. Computer Speech & Language 47, 314–332.

Giri, R., Seltzer, M. L., Droppo, J., Yu, D., April 2015. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, QLD, Australia, pp. 5014–5018.

Guerrero, C., Tryfou, G., Omologo, M., September 2016. Channel selection for distant speech recognition exploiting cepstral distance. In: INTERSPEECH. San Francisco, USA, pp. 1986–1990.

Himawan, I., Motlicek, P., Sridharan, S., Dean, D., Tjondronegoro, D., September 2015. Channel selection in the short-time modulation domain for distant speech recognition. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association. Dresden, Germany, pp. 741–745.

Houtgast, T., Steeneken, H. J., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. The Journal of the Acoustical Society of America 77 (3), 1069–1077.

Ismail, M. R., 2013. A parametric investigation of the acoustical performance of contemporary mosques. Frontiers of Architectural Research 2 (1), 30 – 41.

Ito, N., Makino, T., Araki, S., Nakatani, T., April 2018. Maximum-likelihood online speaker diarization in noisy meetings based on categorical mixture model and probabilistic spatial dictionary. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada, pp. 546–550.

Jiang, Y., Wang, D., Liu, R., Feng, Z., Dec 2014. Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (12), 2112–2121.

Kinoshita, K., Delcroix, M., Gannot, S., Habets, E. A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., et al., 2016. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP Journal on Advances in Signal Processing 2016 (1), 7.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., October 2013. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE, New Paltz, NY, USA, pp. 1–4.

Kitawaki, N., Itoh, K., Honda, M., Kakehi, K., May 1982. Comparison of objective speech quality measures for voiceband codecs. In: ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 7. Paris, France, pp. 1000–1003.

Kitawaki, N., Nagabuchi, H., Itoh, K., 1988. Objective quality evaluation for low-bit-rate speech coding systems. IEEE Journal on Selected Areas in Communications 6 (2), 242–248.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. The annals of mathematical statistics 22 (1), 79–86.

Li, S., Schlieper, R., Peissig, J., 2019. A hybrid method for blind estimation of frequency dependent reverberation time using speech signals. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 211–215.

Madikeri, S., Himawan, I., Motlicek, P., Ferras, M., September 2015. Integrating online i-vector extractor with information bottleneck based speaker diarization system. In: INTERSPEECH. Dresden, Germany.

Malik, H., Farid, H., March 2010. Audio forensics from acoustic reverberation. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Dallas, TX, USA, pp. 1710–1713.

Martínez-González, B., Pardo, J. M., Echeverry-Correa, J. D., San-Segundo, R., 2017. Spatial features selection for unsupervised speaker segmentation and clustering. Expert Systems with Applications 73, 27–42.

Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Andrusenko, A., Podluzhny, I., et al., Oct 2020. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario. Interspeech 2020.
URL http://dx.doi.org/10.21437/Interspeech.2020-1602

Meignier, S., Merlin, T., 2010. LIUM Spkdiarization: An open source toolkit for diarization. In: CMU SPUD Workshop. Dallas, United States.

Oo, Z., Wang, L., Phapatanaburi, K., Iwahashi, M., Nakagawa, S., Dang, J., Jul 2018. Phase and reverberation aware DNN for distant-talking speech enhancement. Multimedia Tools and Applications 77 (14), 18865–18880.

Panday, D., Cordeiro de Amorim, R., Lane, P., 2018. Feature weighting as a tool for unsupervised feature selection. Information Processing Letters 129, 44 – 52.
URL http://www.sciencedirect.com/science/article/pii/S0020019017301618

Parada, P. P., Sharma, D., van Waterschoot, T., Naylor, P. A., 2017. Robust statistical processing of TDOA estimates for distant speaker diarization. In: Signal Processing Conference (EUSIPCO), 2017 25th European. IEEE, pp. 86–90.

Pardo, J., Anguera, X., Wooters, C., 2007a. Speaker diarization for multiple-distant-microphone meetings using several sources of information. IEEE Transactions on Computers 56 (9), 1212–1224.

Pardo, J., Anguera, X., Wooters, C., September 2007b. Speaker diarization for multiple-distant-microphone meetings using several sources of information. IEEE Transactions on Computers 56 (9), 1212–1224.

Peso, P., 2016. Spatial features of reverberant speech: estimation and application to recognition and diarization.

Plapous, C., Marro, C., Scalart, P., Nov 2006. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing 14 (6), 2098–2108.

Sun, L., Du, J., Gao, T., Lu, Y., Tsao, Y., Lee, C., Ryant, N., April 2018. A novel lstm-based speech preprocessor for speaker diarization in realistic mismatch conditions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada, pp. 5234–5238.

Tu, Y.-H., Du, J., Wang, Q., Bao, X., Dai, L.-R., Lee, C.-H., 2017. An information fusion framework with multi-channel feature concatenation and multi-perspective system combination for the deep-learning-based robust recognition of microphone array speech. Computer Speech & Language 46, 517–534.

Vijayasenan, D., Valente, F., Bourlard, H., 2008. Integration of TDOA features in information bottleneck framework for fast speaker diarization. In: Ninth Annual Conference of the International Speech Communication Association.

Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., Snyder, D., Subramanian, A. S., Trmal, J., Yair, B. B., Boeddeker, C., Ni, Z., Fujita, Y., Horiguchi, S., Kanda, N., Yoshioka, T., Ryant, N., 2020. Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings.

Wen, J. Y. C., Habets, E. A. P., Naylor, P. A., 2008. Blind estimation of reverberation time based on the distribution of signal decay rates. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 329–332.

Wolf, M., Nadeu, C., 2014. Channel selection measures for multi-microphone speech recognition. Speech Communication 57, 170 – 180.

Zhang, H., Zhang, R., Nie, F., Li, X., 2019. An efficient framework for unsupervised feature selection. Neurocomputing 366, 194 – 207.
URL http://www.sciencedirect.com/science/article/pii/S092523121930952X

Zhang, Y.-D., Dong, Z., Wang, S.-H., Yu, X., Yao, X., Zhou, Q., Hu, H., Li, M., Jiménez-Mesa, C., Ramirez, J., Martinez, F. J., Gorriz, J. M., 2020. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. Information Fusion 64, 149 – 187.
URL http://www.sciencedirect.com/science/article/pii/S1566253520303183

Zhu, X., Kang, J., Ma, H., 2020. The impact of surface scattering on reverberation time in differently shaped spaces. Applied Sciences 10 (14), 4880.