Directional and Qualitative Feature Classification for Speaker Diarization with Dual Microphone Arrays

Sergei Astapov[a], Dmitriy Popov[b] and Vladimir Kabarov[a]

[a]*International Research Laboratory "Multimodal Biometric and Speech Systems," ITMO University, Kronverksky prospekt 49A, St. Petersburg, 197101, Russian Federation*

[b]*Speech Technology Center, Vyborgskaya naberezhnaya 45E, St. Petersburg, 194044, Russian Federation*

Abstract

Automatic meeting transcription has long been one of the common applications for natural language processing methods. The quality of automatic meeting transcription for the cases of distant speech acquired by a common audio recording device suffers from the negative effects of distant speech signal attenuation, distortion imposed by reverberation and background noise pollution. Automatic meeting transcription mainly involves the tasks of Automatic Speech Recognition (ASR) and speaker diarization. While state-of-the-art approaches to ASR are able to reach decent recognition quality on distant speech, there still exists a lack of prominent speaker diarization methods for the distant speech case. This paper studies a set of directional and qualitative features extracted from a dual microphone array signal and evaluates their applicability to speaker diarization for the noisy distant speech case. These features represent respectively the speaker spatial distribution and the intrinsic signal quality properties. Evaluation of the feature sets is performed on real life data acquired in babble noise conditions by conducting several classification experiments aimed at distinguishing between utterances produced by different conversation participants and between those produced by the background speakers. The study shows that specific sets of features result in satisfying classification accuracy and can be further investigated in experiments combining them with biometric and other types of properties.

Keywords

Meeting transcription, Distant speech processing, Dual microphone arrays, GCC-PHAT, Beamforming, Signal quality features, Artificial neural networks, Classification

1. Introduction

Automatic meeting transcription has long been a common application for Natural Speech Processing (NSP) methods [1, 2, 3, 4]. Automatic meeting transcription and meeting minutes logging is a problem, which mainly employs methods of Automatic Speech Recognition (ASR) and speaker identification and diarization [5]. The available commercial solutions to automatic meeting transcription can be divided into two major groups: ones tending to close-talking speech processing and ones tending to distant speech processing. Close-talking speech processing implies speech acquisition with a close-talking microphone per each speaker. It tends

Proceedings of the 12th Majorov International Conference on Software Engineering and Computer Systems, December 10–11, 2020, Online & Saint Petersburg, Russia

astapov@speechpro.com (S. Astapov); popov-d@speechpro.com (D. Popov); kabarov@speechpro.com (V. Kabarov)

0000-0001-8381-8841 (S. Astapov); 0000-0001-7641-5542 (D. Popov); 0000-0001-6300-9473 (V. Kabarov)

to the scenarios where each speaker has a headset, lapel microphone, or any other type of personal audio recording device. Distant speech processing methods, on the other hand, tend to the scenarios where one or several microphones are situated at a distance to all the attending speakers and, thus, record a mixture of (occasionally overlapping) speech signals incoming from different speakers and background noise incoming from a variety of sources.

Automatic meeting transcription solutions for close-talking speech are quite diverse with several commercial solutions available on the market [6, 7, 1]. Close-talking speech processing poses a significantly less complicated problem for automatic NSP methods if compared to distant speech processing methods. Such, a condition that each attending speaker possesses a personal audio recording device makes the task of speaker identification and diarization almost a trivial one. Voice detection is performed in each independent channel separately and the lack of effects caused by distance as attenuation and interference are negligible. Furthermore, the quality of state-of-the-art methods for ASR perform on par with human perception levels for close-talking speech [8, 7].

Distant speech processing, on the other hand, poses a greater problem for ASR and speaker diarization [9]. The effects of speech signal mixing, signal attenuation due to distance, influence of interference and reverberation, noise pollution — all negatively affect ASR and speaker diarization accuracy. While the state-of-the-art in ASR has reached decent recognition quality on distant speech, there still exists a lack of prominent speaker diarization methods fitted for the task. Part of the recent developments in the field of distant speaker diarization focuses on applying both lexical [10, 11] and acoustic [12, 10] features for model training, usually based on Artificial Neural Networks (ANN) [5]. Application of multichannel sound acquisition devices (microphone arrays) also provides the opportunity to extract spatial features for different sound sources [4, 9, 13]. Along with biometric profile extraction this tends to be a prominent direction of research [14, 15]. Noise pollution remains the main concern for diarization systems, specifically under babble noise produced by speakers not-of-interest situated nearby [16, 17]. Babble noise harshly affects diarization quality based on any type of feature, lexical or acoustic.

This paper considers a set of acoustic features extracted from an audio signal acquired by a dual microphone array. The examined set of features consists of directional features, regarding the spatial disposition of speakers, and signal quality features, regarding the amount of distortion in the signal, closeness to the microphone, estimated Signal to Noise Ratio (SNR) and the T_{60} reverberation time. The applicability of features to the task of distant speaker diarization is evaluated by determining the classification accuracy of speaker utterances belonging either to a target speaker among other active speakers or to the background speakers not-of-interest. Thus the combination of spatial and qualitative acoustic features is aimed at reducing the negative effects of babble noise on diarization quality.

2. Problem Formulation

The considered application of the method described in this paper consists of logging a conversation between two parties of speakers situated on the two opposite sides of a desk or table. Such a scenario rises, for example, in cases of an interview, negotiations, service provision, or any other kind of meeting where such a disposition of parties is appropriate [1, 2]. The audio recording device is placed in the middle of the desk or table between the two parties of speakers. The recording device houses two microphones situated in a straight line parallel to the speaker disposition, i.e., one of the two microphones is directed to one side of the desk and the other — to the opposite side of the desk. The recording device is compact, with a distance between the two microphones measured in several centimeters. The two microphones are sampled synchronously and thus form a dual microphone array.

The task of speaker diarization in the considered scenario consists of estimating the active speaker party per each spoken word or phrase (utterance). If the conversation involves just two active speakers, the distinction of utterances must be made between these two speakers; if any party contains more than one active speaker, an entire party is considered as one speaker [1]. This is allowed in our study as speaker biometric information is not included in the set of examined features. In a more general case employing a greater amount of microphones and involving speaker biometrics it should be possible to expand the dirization task for a more general meeting scenario, where any number of speakers is situated anywhere around the common audio recording device [2]. In this study we primarily address the problem of speaker diarization in babble noise, i.e., a conversation where speakers not-of-interest are present near the conversation area.

3. Applied Features and Methods

The examined features are extracted trough several methods consisting of conventional signal processing and ANN based models. This section addresses these feature extraction methods. Any operation is performed either on the temporal audio signal or its frequency domain representation acquired through the Short-Time Fourier Transform (STFT). The dual channel signal in the time domain is represented by a sequence of observation vectors $\mathbf{x}(t) = [x_1(t), x_2(t)]$ and the STFT representation of the temporal signal is denoted as $\mathbf{X}(t, f) = [X_1(t, f), X_2(t, f)]$, where t is the discrete time instance and f is the STFT frequency band.

3.1. Directional Features

The extracted directional features are based on the Time Difference of Arrival (TDOA) between the two microphones. While the array is placed such, that one microphone is situated closer to the speaker-of-interest that the other, the propagating acoustic waves reach the farthest microphone with a specific delay, compared to the closest microphone. This delay defines the TDOA, which in turn defines the direction to the sound source (speaker). We estimate the TDOA by applying Generalized Cross-Correlation with β-weighted Phase Transform (GCC-βPHAT) [18]. The GCC-βPHAT for a dual microphone array is defined by the following equation:

$$R_{12}(t, \tau) = \Re\left(\frac{X_1(t, f)X_2^*(t, f)}{\left|X_1(t, f)X_2^*(t, f)\right|^\beta} e^{-\mathrm{i}2\pi f\tau}\right), \tag{1}$$

where τ is the time delay between two channels, $\left|X_1(t, f)X_2^*(t, f)\right|^\beta$ is the βPHAT coefficient, $(\cdot)^*$ denotes the complex conjugate, and $\Re(\cdot)$ denotes the real part of a complex number. The range of physically possible time delays between two microphones is defined as $\tau \in [-d/c, d/c]$, where d is the distance between the two microphones and c is the speed of sound in air. The TDOA for time instance t is then defined as the time delay at which GCC-βPHAT reaches its maximal value:

$$\tau_{\text{TDOA}}(t) = \arg\max_{\tau} \left(R_{12}(t, \tau)\right). \tag{2}$$

For a recognized utterance $\text{S} = \{\mathbf{X}(t_1, f), \dots, \mathbf{X}(t_2, f)\}$ ranging in the interval $t \in [t_1, t_2]$ we extract directional features in the following manner. The TDOA values corresponding to this time interval $\boldsymbol{\tau} = \{\tau_{\text{TDOA}}(t_1), \dots, \tau_{\text{TDOA}}(t_2)\}$ are split into two groups of positive and negative values:

$$\begin{aligned} \boldsymbol{\tau}^{(+)} &= \{\tau_{\text{TDOA}}(t) \mid \tau_{\text{TDOA}}(t) > \epsilon, t \in [t_1, t_2]\}, \\ \boldsymbol{\tau}^{(-)} &= \{\tau_{\text{TDOA}}(t) \mid \tau_{\text{TDOA}}(t) < -\epsilon, t \in [t_1, t_2]\}, \end{aligned} \tag{3}$$

where ϵ is a small positive number defining the TDOA ambiguity around zero. As the two speakers or two speaker parties are situated on the opposite sides of the dual microphone array (i.e., along the straight line connecting the two microphones), the TDOA values should be either positive or negative depending on the side of the sound source. As any utterance may include micro-pauses and noise instances, we extract the following directional features based on TDOA:

$$\mathbf{F}_D = \left\{ \frac{\left|\boldsymbol{\tau}^{(+)}\right|}{|\boldsymbol{\tau}|}, \frac{\left|\boldsymbol{\tau}^{(-)}\right|}{|\boldsymbol{\tau}|}, \text{mean}\left(\boldsymbol{\tau}^{(+)}\right), \text{mean}\left(\boldsymbol{\tau}^{(-)}\right), \text{mean}\left(\boldsymbol{\tau}\right) \right\}, \tag{4}$$

where $|\cdot|$ denotes the cardinality of a set. As the set of TDOA values $\boldsymbol{\tau}$ may include both positive and negative values due to noise pollution and pauses (silence), one simple feature of average TDOA along $\boldsymbol{\tau}$ is not sufficient to represent the utterance. It should be addressed that if $\boldsymbol{\tau}^{(+)}$ or $\boldsymbol{\tau}^{(-)}$ have zero cardinality, the mean value is deemed zero:

$$\text{mean}\left(\boldsymbol{\tau}^{(+)}\right) = \begin{cases} \frac{1}{\left|\boldsymbol{\tau}^{(+)}\right|} \sum \boldsymbol{\tau}^{(+)}, & \left|\boldsymbol{\tau}^{(+)}\right| > 0, \\ 0, & \left|\boldsymbol{\tau}^{(+)}\right| = 0. \end{cases} \tag{5}$$

3.2. Qualitative Features

Directional features on their own are quite representative in an ideal case of absence of any background noise. If noise and speech-not-of-interest are present in the signal, and the sources of coherent noise (including background speakers) are situated in the vicinity of the conversation desk or table, the quality of TDOA features may decrease due to masking effects. An example of GCC-βPHAT (discussed in Subsection 3.1) value sequence corresponding to a conversation in noisy conditions with presence of babble noise is presented in Fig. 1. The figure displays a GCC-βPHAT vector $R_{12}(t, \tau)$ from equation (1) for each of the sequence of signal STFT frames. Here the TDOA values corresponding to the actual conversation participants are distributed at the spectral shift index of approximately 90 for one side of the desk and approximately −100 for the other. It can be noticed, that several other distributions exist: at the spectral shift indices of approximately 100 to 200, −200 to −100 and around 0. The first two correspond to babble and other noise and the third one — to uncorrelated diffuse noise and interference. This aspect would not pose a problem in close-talking applications and where ASR

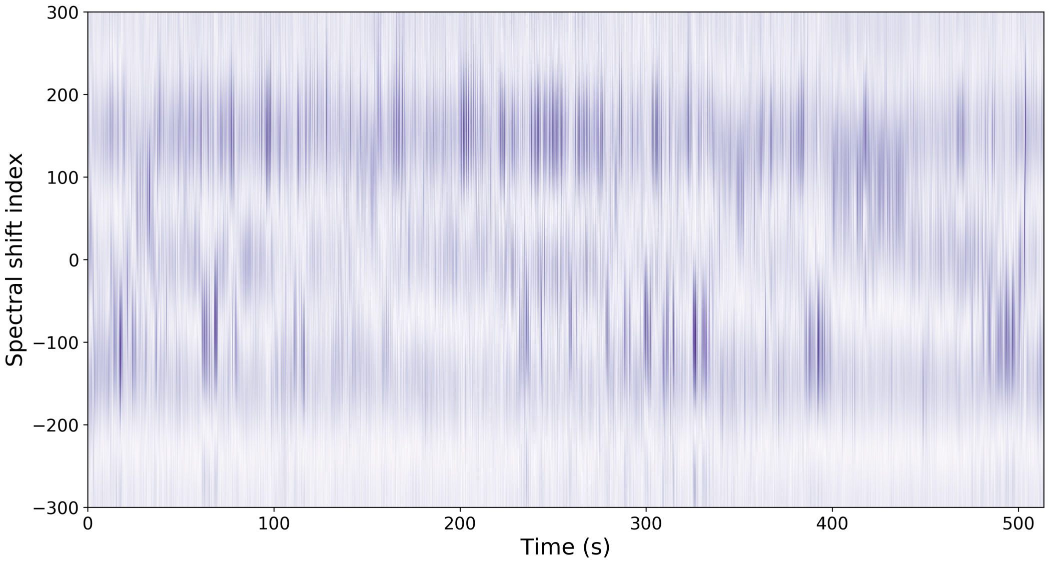

Figure 1: A sequence of GCC-βPHAT frames for a conversation recording. Darker color represents higher GCC values.

would recognize only closest spoken utterances. Unfortunately, in distant speech processing applications this is rarely the case, and phrases spoken in the vicinity of the target conversation are very often recognized by distant speech ASR systems. Explicitly gathering biometric information for all detected speakers would remedy the situation, however, such an approach involves a significant amount of manual manipulations and is not often feasible. We attempt at distinguishing between closest speech of target speakers and farther speech of background speakers by applying several signal quality metrics as features.

The first discussed signal quality metric is the signal Envelope-Variance (EV) [19]. It was originally proposed for the purposes of blind channel selection in multi-microphone systems. It involves calculating the Mel-frequency filterbank energy coefficients of STFT frames, similarly to the process involved in Mel-frequency Cepstral Coefficient (MFCC) calculation [20]. The Mel filterbank coefficients are denoted for an utterance as $\mathrm{S} = \{\mathbf{X}_{\mathrm{Mel}}(t_1, f), \dots, \mathbf{X}_{\mathrm{Mel}}(t_2, f)\}$, where f are log frequencies in the Mel-scale. To remove the short term effects of different electric gains and impulse responses of the microphones from the signal in each channel the mean value is subtracted in the log domain from each sub-band as

$$\hat{X}_i^{\mathrm{Mel}}(t, f) = e^{\log X_i^{\mathrm{Mel}}(t,f) - \mu_{X_i^{\mathrm{Mel}}}(f)}, \tag{6}$$

where $X_i^{\mathrm{Mel}}(t, f)$ is the Mel filterbank coefficient for microphone channel $i \in [1, 2]$, $t \in [t_1, t_2]$; the mean $\mu_{X_i^{\mathrm{Mel}}}(f)$ is estimated by a time average in each sub-band along the whole utterance. After mean normalization, the sequence of Mel filterbank energies is compressed applying a cube root function, and a variance measure is calculated for each sub-band and channel:

$$V_i(f) = \mathrm{var}\left[\hat{X}_i^{\mathrm{Mel}}(t, f)^{\frac{1}{3}}\right]. \tag{7}$$

The cube root compression function is preferred to the conventionally used logarithm, because very small values in the silent portions of the utterance may lead to extremely large negative values after the log operation, which would distort variance estimation. The EV metric is then calculated for each signal channel across all sub-bands as

$$\mathrm{EV}_i = \sum_f \frac{V_i(f)}{\max_i \left(V_i(f)\right)}. \tag{8}$$

EV represents the degree of distortion in a signal; the EV value is higher in a signal channel, where the degree of interference, imposed by distant signal distortion and reverberation, is lower. Thus, the feature vector $\mathrm{EV} = \{\mathrm{EV}_1, \mathrm{EV}_2\}$ not only points to the channel with less distorted speech (speaking party direction), but also gives highlight on the nature of the utterance (either conversation participant, or background talker).

The second and third discussed quality metrics are the Cepstral Distance (CD) and the related Covariance-weighted Cepstral Distance (WCD) [21]. CD is another metric of signal distortion but computed in the cepstral domain, i.e., on the coefficients resulting from MFCC. Cepstrum-based comparisons are equivalent to comparisons of the smoothed log spectra; in this domain the reverberation effect can be viewed as additive. An utterance in the cepstral domain is denoted by $\mathrm{S} = \{\mathbf{c}(t_1, k), \ldots, \mathbf{c}(t_2, k)\}$, where $\mathbf{c}(t, k)$ are the cepstral coefficients and k is the coefficient index (so-called quefrency). For our application we define the distance for each channel i as

$$d_i^{\mathrm{CEP}}(t) = \sum_{k=1}^{p} \left(c_i(t, k) - \bar{c}(t, k)\right)^2, \tag{9}$$

where $c_i(t, k)$ is the cepstral coefficient of $\mathbf{c}(t, k)$ for the i-th channel, $\bar{c}(t, k)$ is the mean cepstral coefficient computed by taking the MFCC from the averaged signal along all channels, and p is the number of cepstral coefficients. The mean coefficients contain the averaged close-talk signal, and the average reverberation component. Let us assume that one microphone signal is better than the others in terms of direct to reverberant ratio. The basic assumption is that such a signal will be characterized by a larger distance from the mean cepstrum. Therefore, from the set of distances $d_i^{\mathrm{CEP}}(t)$ between the mean cepstrum and all the available channels, the least distorted channel can be selected as

$$\hat{M}(t) = \arg\max_i d_i^{\mathrm{CEP}}(t). \tag{10}$$

As an entire utterance can contain some number of coefficients corresponding to noise and silence, we define the CD feature for each channel for the entire utterance as the ratio

$$\mathrm{CD}_i = \frac{\left|\left\{\hat{M}(t) \mid \hat{M}(t) = i\right\}\right|}{\left|\left\{\hat{M}(t)\right\}\right|}, \tag{11}$$

where $|\{\cdot\}|$ denotes the cardinality of a set and $t \in [t_1, t_2]$. The feature vector for a dual channel signal is then $\mathrm{CD} = \{\mathrm{CD}_1, \mathrm{CD}_2\}$. The channel with less distortion will have a higher ratio value and the ratio difference between channels will be greater for close-talking utterances than for the background ones.

The WCD features are obtained in a similar fashion as the CD features, with the only difference being the computation of equation (9). For WCD the $k \times k$ covariance matrix of the cepstral distance vector is computed for each channel:

$$\mathbf{V}_i(t) = \operatorname{cov}\left[c_i(t) - \bar{c}(t)\right], \tag{12}$$

and the covariance weights $w_i(t, k)$ are retrieved as the inverse of every k-th diagonal element v_{kk} of the covariance matrix $\mathbf{V}_i(t)$. The WCD measure is then a weighted Euclidean distance measure, where each individual cepstral component is variance-equalized by the weight:

$$d_i^{\mathrm{WCEP}}(t) = \sum_{k=1}^{p} w_i(t,k)\left(c_i(t,k) - \bar{c}(t,k)\right)^2. \tag{13}$$

The subsequent computations are equivalent to equations (10), (11). The WCD ratio for an entire utterance for all channels is defined as $\mathrm{WCD} = \{\mathrm{WCD}_1, \mathrm{WCD}_2\}$.

Other quality features include Root Mean Square (RMS) energy and common SNR and T_{60} reverberation time estimates. Instant RMS is computed for each channel for each STFT frame as

$$\mathrm{RMS}_i(t) = \sqrt{2\sum_{f=0}^{F_s/2} |X_i(t,f)|^2}, \tag{14}$$

where F_s is the sampling rate and $|X_i(t,f)|$ is the modulus of the complex spectrum. For an entire utterance the RMS energy is computed as the average of instant values $\mathrm{RMS}_i = \mathrm{mean}(\mathrm{RMS}_i(t))$, for $t \in [t_1, t_2]$. And the RMS feature for all channels is $\mathrm{RMS} = \{\mathrm{RMS}_1, \mathrm{RMS}_2\}$. The SNR and T_{60} are estimated by a neural network based voice activity detector (NN VAD) integrated into the the ASR system applied in this study [22]. For every detected and recognized utterance a scalar estimate of the common SNR (dB) and T_{60} (seconds), one for all channels, is provided.

The entire set of qualitative features thus consists of the following features per each utterance:

$$\mathbf{F}_Q = \{\mathrm{EV}, \mathrm{CD}, \mathrm{WCD}, \mathrm{RMS}, \mathrm{SNR}, T_{60}\}. \tag{15}$$

3.3. Beamformed Features

Additionally to extracting features from the raw input signal we study the influence of features extracted from the beamformed signal [18]. For this study we apply two types of simple beamforming algorithms: Delay and Sum Beamforming (DSB) and Differential beamforming (DIF). We intend to examine the influence on the features imposed by steering the dual channel signal in two extreme directions along the linear array (i.e., in the directions to the two participants). For this we apply beamforming in an endfire fashion.

The principle of DSB is expressed in the following equation:

$$X_{\mathrm{DSB}}(t,f,\tau) = \frac{1}{2}\left(X_1(t,f) + X_2(t,f)e^{-\mathrm{i}2\pi f\tau}\right), \tag{16}$$

where $e^{-\mathrm{i}2\pi f\tau}$ is the spectral shift operator at time delay τ, the same as for equation (1). The DIF beamformer, on the other hand, is defined as

$$X_{\mathrm{DIF}}(t,f,\tau)=\frac{1}{2}\left(X_1(t,f)-X_2(t,f)e^{-\mathrm{i}2\pi f\tau}\right). \tag{17}$$

As the frequency response of the DIF beamformer is significantly nonuniform, with the lower frequencies being attenuated in the first half-lobe of the response, we apply an equalizer to the lower frequencies before the cutting frequency of the first lobe $f_c = c/d$ (for the endfire case):

$$H_{eq}(f)=\begin{cases}\left[\sin\left(\pi f\frac{d}{c}\right)\right]^{-1}, & f\le f_c,\\ 1, & f>f_c.\end{cases} \tag{18}$$

The equalizer is then applied to all DIF beamformed frames as $X_{\mathrm{DIF}}(t,f) \longleftarrow H_{eq}(f)X_{\mathrm{DIF}}(t,f)$.

Applying beamforming in an endfire fashion means that the time delay is set to the two extreme values $\tau = \{-d/c, d/c\}$. And so we obtain the two beamformed channels as $\mathbf{X}_{\mathrm{DSB}}(t,f) = [X_{\mathrm{DSB}}(t,f,-d/c), X_{\mathrm{DSB}}(t,f,d/c)]$ and $\mathbf{X}_{\mathrm{DIF}}(t,f) = [X_{\mathrm{DIF}}(t,f,-d/c), X_{\mathrm{DIF}}(t,f,d/c)]$. As beamformed signals lose the initial phase information, they cannot be applied to directional feature extraction. Thus they are applied to extract all the qualitative features excluding SNR and T_{60} as these are provided by NN VAD, which implies only raw signal input. These features are thus: EV, CD, WCD, RMS. The sets of these features extracted on specific beamformed data are denoted as $\mathbf{F}_{\mathrm{DSB}}$ and $\mathbf{F}_{\mathrm{DIF}}$.

3.4. Extraction Procedure Review

The block diagram of the entire feature extraction procedure is presented in Fig. 2. The dual channel signal in the figure denotes the signal interval of one recognized speech segment (utterance). The SNR and T_{60} estimates are retrieved from this signal by the NN VAD [22], as described in Subsection 3.2, and for the extraction of other features it is sufficient to perform all the steps of MFCC separately, while extracting respective features on every intermediate step. Thus, as described in Subsections 3.1, 3.2, after STFT the directional features and RMS are extracted; after computing Mel filterbank energy coefficients the EV features are extracted; and the CD and WCD features are extracted at the last stage of computing cepstral coefficients. The signal is beamformed to additionally extract the features described in Subsection 3.3. This approach reduces the number of MFCC computation instances to just three (one, if beamformers are not involved), which reduces feature extraction computational load.

4. Experimental Evaluation and Results

This section presents the experimental setup for data acquisition, feature classification approaches applied and discusses the feature evaluation results.

4.1. Experimental Setup and Data

The experiments are performed on a signal database of real life recordings of conversations according to the scenario highlighted in Section 2. People were asked to take on their meetings

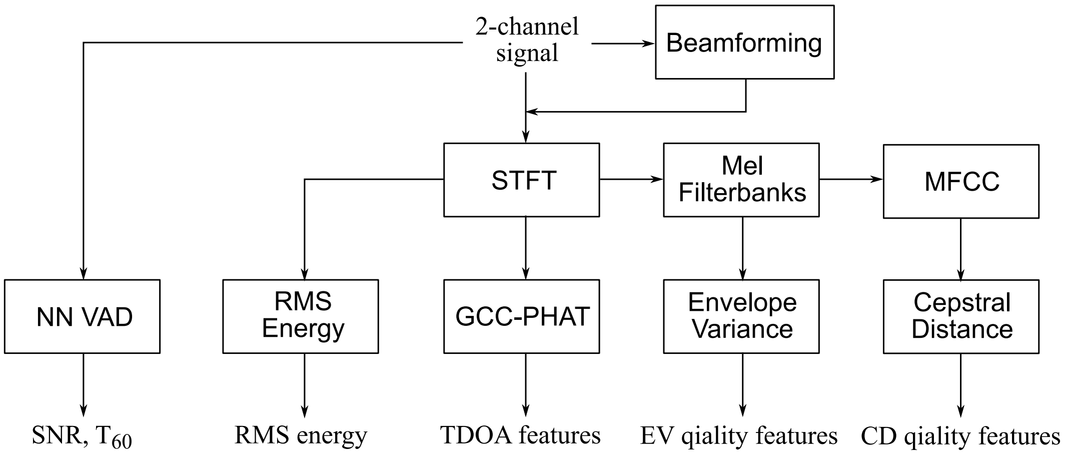

Figure 2: Block diagram of the applied feature extraction procedure.

and discussions at a table with dimensions 2 × 1 meters in an open office space. No physical boundaries were erected around the table, i.e., background speakers were able to move and participate in their daily routine along both sides of the longer side of this table. Several desk phones, a printer and an air conditioning unit were situated in the vicinity of the table. Private discussions (2 participants) and working group meetings (3–6 participants) were taking place at the table with the only restriction being that all the participants had to be seated along both of the longer sides of the table. The dual microphone array with a inter-microphone distance of $d = 0.05$ m was placed at the center of the table perpendicular to the longer side.

The resulting database contains conversations between people of both sexes in a noisy office environment with ample babble noise. The audio signals are acquired in 16 bit PCM WAV files with the sampling rate equal to 16 kHz. The reverberation time measured at the table equals $T_{60} = 450 - 500$ ms. The meeting recordings were manually transcribed with assignment of three distinct classes: person on the left, person on the right side of the table, background speaker. Background speech was transcribed according to typical human hearing capabilities, i.e., distant inaudible speech was not transcribed. Transcription resulted in 115,460 utterances, which were selected in order to reach an approximately equal sample distribution between the three classes.

4.2. Feature Classification

According to Section 3, the feature set comprises a directional feature vector $\mathbf{F}_D$ (4) of length 5, a qualitative feature vector $\mathbf{F}_Q$ (15) of length 10, and two qualitative feature vectors $\mathbf{F}_{\mathrm{DSB}}$, $\mathbf{F}_{\mathrm{DIF}}$ extracted from DSB and DIF beamformed data, both of length 8. The entire feature set then comprises 31 features. The according feature subsets are examined in specific combinations by applying a classification procedure for three variations of target classes:

- 3 class: speaker right (sp1), speaker left (sp2), background speaker (bg);
- 2 class: any of the participant speakers (sp1&2), background speaker (bg);

Table 1
Feature classification results.

Feature set	Classification accuracy (%) for classes		
	sp1, sp2, bg	sp(1&2), bg	sp1, sp2
TDOA	60.3	67.7	76.1
TDOA+Qual	78.2	85.2	89.9
TDOA+Qual+DSB	79.8	87.5	91.8
TDOA+Qual+DIF	79.7	87.6	93.0
All	81.2	87.7	93.2
Selected	80.8	87.8	93.5

- 2 class: speaker right (sp1), speaker left (sp2).

For feature classification an ANN classifier is trained and tested on every examined subset of features. For the first two variations of target classes all utterances from the database are used; for the last variation only participant utterances are used. The classifier is implemented in `tensorflow` and consists of three layers: first dense layer, 80–200 neurons, ReLU activation; second dense layer, 40–100 neurons, ReLU activation, third dense layer, softmax activation. The number of neurons is experimentally selected to benefit classification of different feature subsets. `EarlyStopping` and `ReduceLROnPlateau` are applied during training for monitoring the validation loss.

Additionally feature selection is performed by applying Recursive Feature Elimination with Cross-Validation (RFECV) from the `sklearn` package while iteratively training a classifier of type `GradientBoostingClassifier` on combinations of features with elimination factor equal to 1. As a result the selected feature set excludes 10 features without significant classification accuracy loss.

4.3. Evaluation Results

The results of feature subset evaluation are presented in Table 1. It is evident from the table that directional (TDOA) $\mathbf{F}_D$ features alone cannot distinguish between the 3 classes defined in Subsection 4.2 beyond the margin of random choice. Futhermore, for the first two classification problems the recall of class 3 (bg) is the lowest, which means that applying just TDOA features does not provide near vs far speaker separation. Applying qualitative features, on the other hand, significantly improves classification quality for all three classification tasks. The application of beamformed data improves the quality just slightly and, therefore, may be superfluous for limited resource solutions. Classification on the selected feature set, which incorporates 21 of 31 features (reduction by 32%), is on par with the accuracy over the entire feature set. This implies that single features from all regarded subsets are redundant for the classification task at hand. However, this may be the case for this specific dataset.

Generally, the discussed directional and qualitative features seem to be applicable for the task of distinguishing between the participating speakers and background speakers. Qualitative features significantly improve classification accuracy. The established feature set can be considered for further investigation in combination with biometric and other types of features.

5. Conclusion

The paper regarded a set of directional and qualitative signal features for the task of speaker diarization. The discussed feature set is proven to be applicable to the task of speaker utterance classification for the distant speech processing case in office and babble noise conditions. The feature set can be considered for further investigation along in combination with biometric and other types of features.

Acknowledgments

This research was financially supported by the Foundation NTI (Contract 20/18gr,
ID 0000000007418QR20002) and by the Government of the Russian Federation (Grant 08-08).

References

[1] A. Nedoluzhko, O. Bojar, Towards automatic minuting of the meetings, in: 19th Conference ITAT 2019: Slovenskocesky NLP workshop (SloNLP 2019), CreateSpace Independent Publishing Platform, 2019, pp. 112–119.

[2] C. Bokhove, C. Downey, Automated generation of ”good enough” transcripts as a first step to transcription of audio-recorded data, Methodological Innovations 11 (2018).

[3] T. Hain, J. Dines, G. Garau, M. Karafiát, D. Moore, V. Wan, R. Ordelman, S. Renals, Transcription of conference room meetings: an investigation, in: INTERSPEECH, 2005.

[4] T. Yoshioka, Z. Chen, D. Dimitriadis, W. Hinthorn, X. Huang, A. Stolcke, M. Zeng, Meeting transcription using virtual microphone arrays, in: ArXiv, 2019.

[5] H. Aronowitz, W. Zhu, M. Suzuki, G. Kurata, R. Hoory, New advances in speaker diarization, in: INTERSPEECH, 2020.

[6] Y. Huang, Y. Gong, Acoustic model adaptation for presentation transcription and intelligent meeting assistant systems, in: ICASSP 2020, IEEE, 2020.

[7] F. Filippidou, L. Moussiades, Alpha benchmarking of IBM, Google and Wit automatic speech recognition systems, in: I. Maglogiannis, L. Iliadis, E. Pimenidis (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, Cham, 2020, pp. 73–82.

[8] A. Stolcke, J. Droppo, Comparing human and machine errors in conversational speech transcription, ArXiv (2017).

[9] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang, A. Hurvitz, L. Jiang, S. Koubi, E. Krupka, I. Leichter, C. Liu, P. Parthasarathy, A. Vinnikov, L. Wu, X. Xiao, W. Xiong, H. Wang, Z. Wang, J. Zhang, Y. Zhao, T. Zhou, Advances in online audio-visual meeting transcription, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 276–283.

[10] S. Horiguchi, Y. Fujita, K. Nagamatsu, Utterance-wise meeting transcription system using asynchronous distributed microphones, in: INTERSPEECH, 2020.

[11] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, S. Watanabe, Auxiliary interference speaker loss for target-speaker speech recognition, in: INTERSPEECH, 2019.

[12] S. Astapov, D. Popov, V. Kabarov, Directional clustering with polyharmonic phase estimation for enhanced speaker localization, in: A. Karpov, R. Potapova (Eds.), Speech and Computer, Springer International Publishing, Cham, 2020, pp. 45–56.

[13] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, T. Nakatani, Speaker-aware neural network based beamformer for speaker extraction in speech mixtures, in: INTERSPEECH, 2017.

[14] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko, Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario, in: Interspeech 2020, 2020, pp. 274–278.

[15] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with self-attention, 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2019) 296–303.

[16] H. R. Muckenhirn, I. L. Moreno, J. Hershey, K. Wilson, P. Sridhar, Q. Wang, R. A. Saurous, R. Weiss, Y. Jia, Z. Wu, Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking, in: ICASSP 2019, 2018.

[17] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, T. Nakatani, Single channel target speaker extraction and recognition with speaker beam, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5554–5558.

[18] M. Brandstein, D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Digital Signal Processing - Springer-Verlag, Springer, 2001.

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

[20] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (1980) 357–366.

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

[22] G. Lavrentyeva, M. Volkova, A. Avdeeva, S. Novoselov, A. Gorlanov, T. Andzhukaev, A. Ivanov, A. Kozlov, Blind speech signal quality estimation for speaker verification systems, in: Proc. Interspeech 2020, 2020, pp. 1535–1539.