Improved Location Features for Meeting Speaker Diarization

Scott Otterson

University of Washington, Seattle, USA

scotto@u.washington.edu

Abstract

This paper proposes several improvements to the correlation-based location features recently used in meeting speaker diarization. A speech-specific alternative to the generalized cross correlation phase transform (GCC-PHAT) algorithm is tested and shown to provide equal or better results without noise reduction or continuity-enforcing smoothing. The limitations of a single correlation reference waveform are discussed, and it is shown how a multi-band energy ratio feature can help overcome them, yielding significantly improved performance. An all-pairs correlation is also proposed, and when combined with energy ratios, it also improves upon the baseline system. However, the best combination is the baseline correlation features with energy ratios.

Index Terms: speaker identification, diarization, localization

1. Introduction

The goal of meeting diarization is to extract short segments of conversational speech and to associate them with the correct speaker. Most recent state-of-the-art work has been done for the NIST meeting recognition evaluation [1], where several labs competed to transcribe multi-microphone meetings. In the meeting recording scenario (and in the evaluations), there are different possible microphone configurations, including use of a single microphone, multiple microphones with ad hoc placement, and microphone arrays. The paper concentrates on the NIST multiple-distant-microphone test condition (MDM), in which a variable (meeting-dependent) number of microphones are placed at unknown distances from the speakers in an unknown (ad hoc) configuration.

Most meeting diarization systems use traditional speaker identification features, such as mel-warped cepstral coefficients (MFCC’s) but some systems [2] have also used speaker azimuths estimated by beamforming signals from microphone arrays in known locations. The innovation of the most successful MDM system in the NIST 2006 competition [3] was to use “location features,” the set of correlation lags computed between microphones located at unknown locations.

In [3], one of the (noise-reduced) signals coming from the distant microphones is chosen as a reference and then a location feature is computed from the cross-correlations between that reference and each of the remaining distant microphone waveforms. For each correlation pair, four candidate correlation peaks are chosen – usually the four highest – and then a dynamic programming algorithm running across all frames in a meeting selects the best peak for each pair of each frame. If there are N channels, then the correlation feature vector for the i-th frame is composed of the N-1 lags at the best peaks. On several NIST corpora, these features demonstrably improved speaker diarization over using MFCC’s alone.

This paper propose several improvements to the correlation-based location features used in [3]. The key new ideas are to use all microphone pairs for computing correlation features (in combination with dimension reduction via principal components analysis), and to combine correlation features with an all-pairs vector of multi-band energy ratios for increased robustness. Also tested is whether or not a speech-specific inverse filtering and Hilbert envelope correlation calculation is better than the general purpose GCC-PHAT.

The following section summarizes the limitations of location features derived from single-reference channel correlations, describes an improved correlation algorithm, and then presents experimental results suggesting how multi-band energy ratios can increase diarization accuracy.

2. Limits of Single-Reference Correlations

In meetings recorded with ad hoc microphone placement, the microphones are generally widely spaced, meaning that the maximum time-of-flight difference between them can be several times longer than the fundamental period of speech, T0. When coherence is calculated over a time extent long enough to capture the time-of-flight difference, the analysis window will contain several pitch periods. This does not usually cause a problem for sentence-length correlation analysis windows, in which T0 and its harmonics can vary widely enough to produce a single strong coherence peak. However, the goal here is to segment utterances as brief as 0.5s to capture backchannel utterances (“yeah”, “uh-huh”). When coherences are computed over this short analysis window, speech can appear nearly stationary and periodic; for a single location, there will be several coherence peaks of nearly equal magnitude, causing the location feature to be ambiguous in a manner similar to the ambiguous peaks often encountered in correlation-based pitch detectors [4]. In addition, widely spaced microphones may be subject to significantly different room reverberation effects, making a pair of waveforms more dissimilar than with closely spaced array microphones, and resulting in weaker true-delay correlation peaks.

In [3], the problem of multiple and weak peaks due to spatial aliasing and room reverberation was reduced by storing more than one correlation peak for each frame of a pair and then by selecting the most likely peak trajectory among them with a continuity-enforcing dynamic programming algorithm. Since microphone positions are unknown, the dynamic programming algorithm must operate over both high quality correlations coming from closely spaced microphones and low quality correlations coming from widely spaced microphones.

To assess the potential impact of low quality correlations, a series of speaker ID cheating experiments were performed in which known speaker segmentation information is used to extract training data for each speaker. For each correlation pair, a Gaussian mixture model (GMM) is trained for each speaker – the 1-dimensional feature being the lag at maximum correlation. For each pair, a maximum likelihood speaker identification is then made and speaker diarization error is calculated. The result provides an indication of the suitability of a microphone
pair for use in an unsupervised speaker clustering system.

Figure 1: Minimum, median and maximum diarization error using single cross-correlation pair but known (oracle) speaker segmentations in training. Diarization error is shown as a function of the different meetings in the RT05S corpus.

Figure 2: Speaker ID F-number Detector Accuracy (intensity) for microphone pair (horizontal) and each speaker (vertical), based on the correlation for that microphone pair. NIST RT05S corpus, meeting NIST_20050427_0939.

In this section, methods to improve the correlation features in [3] are described, and it is argued that the addition of a new energy ratio feature can overcome the shortcomings inherent in any correlation feature. Combined, the various changes lead to a high dimensional feature vector, which is more effectively used in diarization when transformed to a smaller dimension. Below, is a description of the raw feature vectors, followed by the dimension reduction strategy.

3. Proposed Feature Improvements

Cross-correlation features were implemented in which a feature vector is composed of the peak magnitude lags from all channel pairs. In addition, there are three differences in the signal processing strategy for these features: i) use an alternative method for computing correlations, ii) eliminate the noise reduction step, and iii) eliminate the dynamic programming peak picking step. These differences are described and motivated below.

The GCC-PHAT time delay estimator [6] is popular for use in acoustic localization in reverberant environments. Although it is general purpose and not tailored to speech, it is employed by many researchers concentrating on localizing speech, including [3]. In GCC-PHAT, the cross-correlation is calculated in the frequency domain, after the cross-spectra whitening. However, other researchers [7] have found that, for speech, GCC-PHAT is outperformed by speech-specific algorithms which correlate the Hilbert envelopes of inverse linear-prediction-filtered waveforms. Inverse filtering performs roughly the same function as GCC-PHAT whitening; frequencies are still effectively being divided by their magnitudes, but with inverse filtering, the denominator is smoothed by the spectral envelope of speech, thus avoiding noise pumping in spectral zeros. Viewed in the time domain, the Hilbert envelope calculation, which smooths out easily aliased periodic information in the glottal transient oscillations, is similarly speech-specific. Finally, the delay features are the lags at the peak of standard correlations between the Hilbert envelopes of the inverse filtered channel waveforms.

In addition, no noise reduction is applied before the correlation calculation, as it was found that per-channel noise reduction reduced the pair similarity needed for a good correlation peak. Several researchers, for example, [8, 9], have found that Weiner filter noise reduction before correlation improves diarization performance. We speculate that our speech-specific correlation feature processing is itself a type of noise reduction, and therefore does not benefit from this step.

Rather than using several hypothesized correlation lags from each channel pair followed by the continuity-enforcing dynamic programming algorithm employed in [3], we use a single correlation lag from all channel pairs, followed by dimensionality reduction. The dimensionality reduction, to some extent, improves the reliability of the correlation lags by effectively choosing the most informative channel pairs in the vector.
3.2. All-pairs multi-band energy ratios

Given the problem of widely spaced microphones in unknown positions – unanticipatable reverberation and spatial aliasing – some degradation of correlation features is unavoidable. However, in situations where correlation features fail, an energy ratio feature often performs well.

Sound intensity decreases with distance, and in addition, room reverberation imposes location and frequency-dependent zeros over the sound spectrum. For cross-correlations, these are both problems because they make channel waveforms dissimilar, lowering correlation peaks. For energy ratios, however, this dissimilarity is a speaker location feature.

For this paper, each channel was fed into a filter bank with linearly spaced bandpass filters of width 1250 Hz, centered about 875, 2125 and 3375 Hz. Then, for each pair, the log energy ratio is calculated in each band, yielding a \(3 \times N_{\text{pairs}}\) feature vector, where \(N_{\text{pairs}}\) is the number of channel pairs in the meeting. The filterbank parameters were chosen by a grid search optimizing for cheating experiment diarization accuracy (as in Section 2). The overall bandwidth covers regions of the speech spectrum where there is significant path-length attenuation for distances typical of meeting room environments. The number of bands chosen by the search is a trade-off between a desirably small feature vector dimension and the goal of capturing frequency-dependent room reverberation differences. Although noise reduction did not help with the proposed cross-correlation features, it was found to improve results as a pre-processing step to the energy ratio calculation.

Figure 3 illustrates a meeting where energy ratios can provide improved results for speakers that are difficult to identify with cross-correlation features. The figure shows confusion images, where pixel intensity is proportional to the fraction of cases where an active speaker (vertical axis) is labeled as one of the possible set of speakers in the meeting (horizontal axis). In the cheating experiment speakers are identified with GMM’s trained in a supervised manner on the two different types of location features. As can be seen in the figure, a correlation feature classifier has a higher degree of confusability among candidate speakers.

Others researchers have used energy ratios for diarization [10, 11] however, these were with personal microphones, known to be close to the mouths of the speakers – an easier case – and were not multi-band, meaning that the energy ratios did not characterize detailed frequency and position-dependent room reverberation.

3.3. Dimensionality Reduction

High dimensionality is a major disadvantage of all-pairs location features in multiple frequency ranges – in some meetings in the NIST corpora, the combined energy and cross-correlation location feature vector dimension is 480. The high dimension has high computational cost, which can cause problems with clustering, so both accuracy and efficiency benefit from dimensionality reduction.

The particular approach to dimensionality reduction used here is principal components analysis (PCA) since the meeting-dependent nature of the all-pairs features makes supervised learning of a discriminant transform impossible. Within this framework, we investigate transforms that operate jointly on the concatenated vector of correlation lags and energy ratios, as well as on these vectors independently. The final dimension of the vector is determined empirically.

4. Experimental Paradigm

Features for this paper were tested on data from the NIST Rich Meeting Transcription Project 2004 dev test and 2005 eval data sets [5, 1]. In these meetings, held at five locations in conventional conference rooms, 4 to 10 participants were recorded with 1 to 16 distant, omni-directional microphones. Sound was acquired at 16 bits and 16 KHz. No information about microphone or speaker location is available.

The baseline for comparison in evaluating our features is the system described in [3]. The authors made the feature extraction and clustering software available for this work. The diarization algorithm uses an agglomerative clustering scheme with a Bayesian information criterion (BIC) stopping threshold, and which employs a hidden Markov model (HMM) to enforce segmentation continuity, has been shown to work well in estimating the correct number of speakers and their times. For this work, an extra HMM stream was added for the energy ratio features. In addition, the correlation calculation software used in [3] was modified so that the effect of the dynamic programming smoother could be switched off. This code also does a delay-sum beamforming of the distant microphone channels; MFCC speaker ID features used in this work were derived from this summed channel.

Speaker diarization results were scored against speech recognizer forced alignments using the NIST corpora diarization scoring software.

5. Results

Diarization error for different combinations of features is shown in Table 1 for meetings in the RT04S and RT05SE corpora which have more than one distant microphone. The “best” and “worst” columns are the best and worst scores for individual meetings, while the “tot” column is the overall score for the corpus. The location feature experiments performed were:

- **0**: baseline: the system used in [3].
- **XC1**: all-pairs, inverse filtered, Hilbert transformed correlations, as in Section 3.1, PCA-reduced to 3 dimensions (1 component GMM)
- **XC2**: single-reference correlation, signal processing as in XC1 (1-component GMM)
- **XC3**: single-reference correlation, signal processing as in XC1, 3 PCA components (1-component GMM)
- **ER1**: baseline correlations as in 0, with energy ratios, 2 PCA components, separate HMM stream (15 component GMM)
- **ER2**: all-pairs correlations concatenated onto the energy ratios and PCA-reduced to two dimensions (3 component GMM)
Table 1: Diarization Error vs. Features and Configuration

<table>
<thead>
<tr>
<th>Expt.</th>
<th>Diarization Error</th>
<th>r04s</th>
<th>worst</th>
<th>best</th>
<th>tot</th>
<th>r05se</th>
<th>worst</th>
<th>best</th>
<th>tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>15.5</td>
<td>9.2</td>
<td>12.3</td>
<td></td>
<td>37.2</td>
<td>2.0</td>
<td>14.8</td>
</tr>
<tr>
<td>XC1</td>
<td></td>
<td></td>
<td>20.7</td>
<td>9.1</td>
<td>12.7</td>
<td></td>
<td>37.2</td>
<td>2.5</td>
<td>14.5</td>
</tr>
<tr>
<td>XC2</td>
<td></td>
<td></td>
<td>26.2</td>
<td>5.1</td>
<td>12.5</td>
<td></td>
<td>36.4</td>
<td>2.85</td>
<td>13.9</td>
</tr>
<tr>
<td>XC3</td>
<td></td>
<td></td>
<td>17.9</td>
<td>7.7</td>
<td>11.1</td>
<td></td>
<td>26.4</td>
<td>2.2</td>
<td>14.8</td>
</tr>
<tr>
<td>ER1</td>
<td></td>
<td></td>
<td>16.1</td>
<td>5.51</td>
<td>11.1</td>
<td></td>
<td>35.6</td>
<td>2.12</td>
<td>10.8</td>
</tr>
<tr>
<td>ER2</td>
<td></td>
<td></td>
<td>12.3</td>
<td>5.4</td>
<td>10.3</td>
<td></td>
<td>37.3</td>
<td>1.5</td>
<td>16.3</td>
</tr>
<tr>
<td>ER3</td>
<td></td>
<td></td>
<td>25.7</td>
<td>5.3</td>
<td>11.9</td>
<td></td>
<td>41.7</td>
<td>3.0</td>
<td>13.9</td>
</tr>
<tr>
<td>ER4</td>
<td></td>
<td></td>
<td>51.6</td>
<td>6.5</td>
<td>11.4</td>
<td></td>
<td>35.4</td>
<td>2.8</td>
<td>13.4</td>
</tr>
</tbody>
</table>

ER3: all-pairs correlations (2-PCA dimensions) and energy ratios (2-PCA dimensions), each on an independent HMM stream (1 and 5 component GMM’s for correlation and energy ratios’ respectively)

ER4: baseline correlations (2-PCA dimensions) and energy ratios (2-PCA dimensions), each on an independent stream (1 and 5 component GMM’s for correlation and energy ratios’ respectively)

For each row of the table, HMM stream weights were set to optimize performance; they were set to the same value for each corpus – 0.75 to 0.9 for the MFCC channel and 0.1 to 0.25 for the sum of the remaining channels.

Comparing XC1 to the baseline, we find gains for the RT05 test set, but a small loss for the other. Since there are several differences in processing besides the all-pairs approach (i.e. noise reduction, dynamic programming smoothing, correlation via GCC-PHAT), another feature set, XC2, was created using the same processing but with correlations against the single reference channel chosen in experiment 0. Performance is roughly the same as the baseline, with a slight gain on RT05SE and a smaller loss on RT04S. Without noise reduction or dynamic programming smoothing, the proposed correlation feature signal processing has performance equivalent to the baseline system. The XC3 row of Table 1 shows the effect of dimension reduction. In this case, the baseline is matched or significantly bettered when the features of XC2 are dimension-reduced to three PCA components. While the proposed correlation signal processing and dimension reduction appear to have merit, the benefit of calculating correlations across all pairs is not clear.

Comparing ER1 to the baseline 0, it is clear that energy features lead to improved performance, but there are several questions as to the best way to combine the features. In cheating experiments discussed in Sections 2 and 3, it was found that the optimal number of GMM mixtures was quite different for energy ratios and correlations. Therefore, it was thought that performance would be better if these features were separately dimension-reduced and then fed into independent HMM streams. Comparing ER2, in which the all-pairs correlation and energy ratio features are concatenated, with ER3, in which they are fed to separate streams, it appears that better results can be expected from independent streams.

Rows ER3 and ER4 illustrate the effects of correlation feature dimension reduction when energy ratios are fed to separate streams. Comparing with ER1 it appears that the best correlation and energy ratio feature combination has the baseline correlation features without dimension reduction on one stream and energy ratios on another.

6. Conclusions

This paper has shown several ways to improve speaker diarization location features. First, multi-band energy ratios were shown to significantly improve diarization performance. In a speaker clustering system such as [3], the best results are obtained when correlation features are kept in separate, independent HMM streams instead of concatenated, probably because their characteristics are best modeled by GMM’s with different feature dimensions and number of GMM components.

The all-pairs correlation features proposed here had several anticipated advantages, and indeed, they yielded better performance than the baseline system in some meetings. However, the best performance comes from a single-reference GCC-PHAT correlation features combined with energy ratios.

7. References


