Study on Integration of Speaker Diarization with Speaker Adaptive Speech Recognition for Broadcast Transcription

Jan Silovsky, Petr Cerva, Jindrich Zdansky and Jan Nouza

Institute of Information Technology and Electronics, Faculty of Mechatronics, Technical University of Liberec, Liberec, Czech Republic
{jan.silovsky,petr.cerva,jindrich.zdansky,jan.nouza}@tul.cz

Abstract
In this paper we study a close incorporation of speaker diarization with speaker adaptive speech recognition in our broadcast transcription system. We provide our motivation for utilization of speech transcripts in the diarization process and analyze the effect it yields in terms of diarization performance or computational cost. Further, speaker adaptation performed according to various scenarios of speaker segmentation and diarization of an audio stream is evaluated. For better insight, the limit performance is evaluated substituting most of the components of the system by the oracle ones.

Index Terms: Speaker diarization, i-vectors, speaker adaptation, CMLLR, broadcast transcription

1. Introduction
Unsupervised speaker adaptation is considered a standard part of most of complex speech recognition systems, e.g. those developed by various sites participating in the EARS, GALE, AMI, CHIL programs and others. Usually speaker adaptation techniques require a knowledge of a preliminary transcript of the adaptation data and hence multiple recognition passes are performed. Further, a knowledge about attribution of temporal segments of a continuous stream to appropriate originating speakers is required. Speaker diarization is hence a useful preprocessing step with respect to speaker adaptation as it aims to determine the number of speakers as well as their occurrence in the given audio stream [1, 2].

On the other side, speech recognition could be a useful preprocessing step for speaker diarization [3]. In this paper we aim to analyze the effect that a knowledge of speech transcripts yields to the speaker diarization process and the effect speaker diarization in return yields to speaker adaptive (SA) speech recognition. To gain better insight, various components of the system are substituted by oracle components. By which we refer to utilization of human-produced reference annotations (both speaker segmentation and speech transcripts) instead of automatic outputs provided by the components.

In our system, we make use of information about word boundaries and we also take advantage of classification of various non-speech events as provided by our automatic speech recognition (ASR) system. This allows the diarization system to neglect various noises produced by speakers (breathing, various hesitation sounds, cough, lip-smack, etc.) that carry no speaker-specific information (considering representation of the signal by cepstral features) and thus harm the representation of clusters. Moreover, distribution of the cepstral features corresponding to these sounds differs notably from the distribution corresponding to speech regions and hence a false change-point is often detected due to high values of segmentation test statistics. However, at the same time, these sounds should not break the continuity of segments uttered by a single speaker and should thus not be simply dismissed. Although linguistic information contained in transcripts could provide further cues for discrimination between speakers [4], we avoid its use as it would make the speaker diarization system language dependent. To show how the transcripts are treated, in the next section we present a detail description of our speaker diarization system.

2. Speaker diarization system
The traditional framework of a speaker diarization system consists of three basic modules [5] that perform tasks of speech activity detection (SAD), speaker change-point detection and speaker clustering. In our system, availability of speech transcripts allows us to completely dismiss the output of the standard SAD module (e.g. energy or model based).

2.1. Utilization of speech transcripts
The non-speech events are categorized into two classes depending on whether they were produced by speakers (breathing, various hesitation sounds, cough, lip-smack, etc.) or they are artificial (e.g. music, background noise). Let us denote these classes $C_s$ and $C_n$ respectively. In addition, all lexicon entries (words) are comprised in the class $C_s$. Further, let $p = \{p_1, \ldots, p_M\}$ be a sequence of times corresponding to start position of transcription elements $w_i (i = 1, \ldots, M)$ in a transcript produced by the ASR module. For a stream represented by a sequence of $T$ feature vectors, we then introduce two binary sequences $a_h(t)$ and $a_n(t)$ where $t = 1, \ldots, T$. Values of $a_h(t)$ and $a_n(t)$ are given as

\[ a_h(t) = \begin{cases} 1 & \text{if } w_i \in \{C_s, C_h\} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad a_n(t) = \begin{cases} 1 & \text{if } w_i \in C_s \\ 0 & \text{otherwise} \end{cases} \]

where $p_i < t < p_{i+1}$ and $i = 1, \ldots, M$.

Let us highlight that we do not use the information about speech activity (as determined by the classification of transcript elements) to split the signal but all frames (accompanied by the $a_h(t)$ and $a_n(t)$ values) of the stream are passed to the speaker segmentation step. Hence, the speaker homogeneous segments interleaved by non-speech intervals are not broken regardless of the kind and duration of these intervals. Smoothing of short speech or non-speech intervals that break the fluency of the output is postponed at the end of the diarization process.

The fact that the diarization process does not rely on lexical content of transcripts brings the advantage of lower dependency on language of the ASR module.
2.2. Speaker segmentation

Let $\mathcal{O} = \{o_1, \ldots, o_T\}$ be a sequence of $d$-dimensional feature vectors (frames) representing an audio stream. The sequence is traversed by a sliding variable-length window and a frame within the window is claimed to be a change-point based upon a test of hypotheses. Hypotheses testing requires derivation of proper test statistics. We use the test statistic derived based on the maximum likelihood approach [6] as defined as

$$\Lambda(t) = \alpha \sqrt{n \log |\Sigma| - n_1 \log |\Sigma_1| - n_2 \log |\Sigma_2|} - \beta$$

(2)

where $n_1$ and $n_2$ denote the number of effective frames in the left and right partition of the analyzed window as split by a hypothesized change-point at the time $t$, $\Sigma_1$ and $\Sigma_2$ are full covariance matrices corresponding to these partitions and $\Sigma$ is a full covariance matrix of all effective frames in the analyzed window. Finally, the remaining terms in (2) are defined as:

$$\alpha = (2 \log \log n)^{1/2},$$

(3)

$$\beta = 2 \log \log n + d \log \log n - \log \Gamma(d).$$

(4)

The information about the start position of transcript elements is used as the constraining condition for choice of change-point candidates. Hence a single change-point candidate within the analysis window is found according to the equation

$$\hat{t} = \arg \max_{t \in [1, n-d]} R(t), \quad t \in p.$$  

(5)

Let us remark that as we are using automatic transcripts, this does not mean that a change-point cannot be detected within a word actually uttered by a speaker, but the chance of such detection is significantly decreased. When $\Lambda(\hat{t})$ exceeds a given decision threshold $\theta$, the change-point at the time $\hat{t}$ is confirmed. The decision threshold is set so as to prefer over-segmentation to miss of change-points (the threshold is lower than the optimal threshold found by the training algorithm described in [6]). This is preferable as false detections may be eliminated in the clustering stage while missed change-points are unrecoverable.

The covariance matrix (see eq. (2)) for a partition of the stream ranging from the time $t_1$ to $t_2$ is computed based on the first-and second-order statistics as follows

$$\Sigma = \frac{1}{n} \Delta S - \frac{1}{n^2} \Delta F (\Delta F)^T$$

(6)

where $n = \sum_{t=t_1}^{t_2} a_n(t)$, $\Delta F = \sum_{t=t_1}^{t_2} a_n(t) \Delta f$ and $\Delta S = \sum_{t=t_1}^{t_2} a_n(t) \Delta s^o_t$. In practice, efficient computation of the $\Delta F$ and $\Delta S$ is achieved by differentiation of first-and second-order statistics iteratively accumulated in a circular buffer [6].

2.3. Speaker clustering

Our clustering module uses bottom-up clustering. More specifically, we employ the two-stage clustering scenario described in [7]. At the first pre-clustering stage, the similarity of clusters is measured via the standard criterion based on the Bayesian Information Criterion (BIC) difference. At the second stage, clusters are represented by i-vectors and their similarity is measured by their cosine distance [8].

In the i-vector concept, a simple factor analysis model is employed to extract a fixed-and low-dimensional representation of a segment of variable length in the total variability space (TVS) [8]. A projection from a sequence of feature vectors $O = \{o_1, \ldots, o_T\}$ representing an audio segment to TVS is provided by computation of a Maximum A Posteriori (MAP) point estimate of the so called i-vector $x$ based on the zero-and first-order sufficient statistics gathered employing a Gaussian Mixture Model (GMM) as follows [9]:

$$x = \left(1 + \sum_{c=1}^{C} T_c \Sigma_c^{-1} (T_c)^{-1} \sum_{c=1}^{C} T_c \Sigma_c^{-1} \hat{F}_c (m_c) \right)^{-1} \sum_{c=1}^{C} T_c \Sigma_c^{-1} \hat{F}_c (m_c)$$

(7)

where $T$ is a low-rank rectangular matrix representing the total variability space which can be decomposed into $T_c$ blocks so that $T = \left[T_1 \cdots T_C\right]^T$. $m_c$ and $\Sigma_c$ are a mean vector and a diagonal covariance matrix corresponding to the $c$-th component of the GMM (having $C$ components in total) respectively. Finally, the zero-and (centralized) first-order statistics are computed respectively as follows

$$N_c = \sum_{t} \gamma_c(t) a_n(t)$$

(8)

$$\hat{F}_c = \sum_{t} \gamma_c(t) a_n(t) (o_t - m_c)$$

(9)

where $\gamma_c(t)$ is the posterior probability of the event that feature vector $o_t$ is accounted for by the $c$-th component of the GMM. This way of treatment of transcripts thus results in a form of a frame level paralelization [10].

Having the i-vector representation of segments (or clusters) $g_1$ and $g_2$ by i-vectors $x_1$ and $x_2$ respectively, we can assess their similarity simply using the cosine distance as

$$d(g_1, g_2) = \frac{x_1^T x_2}{|x_1||x_2|}.$$  

(10)

The higher the cosine distance is, the more likely both segments originate from the same speaker. In the clustering process, the two clusters with the highest cosine distance score are merged together and if a maximum value of the score for any pair of clusters drops under a given threshold, the stopping condition of the process is met.

Because Eq. (10) can be applied only for a pair of i-vectors, the i-vector representing a certain cluster is simply computed as an average of all i-vectors corresponding to the initial segments assigned to the given cluster so far [11]. The Linear Discriminant Analysis (LDA) is employed to cope with the nuisance intra-speaker variability.

2.4. Diarization output smoothing

The aim of the smoothing is to discard either short speech or non-speech intervals that harm the fluency of the output as high granularity of the output is very obtrusive to users.

While the segmentation and clustering modules use $a_n(t)$ values to distinguish between the speech and non-speech frames, here $a_n(t)$ values are taken into account. Speech and non-speech segments shorter than 0.25 s and 1.0 s respectively are discarded (in the order as listed).

3. Speech recognition system

We use our Czech LVCSR system to obtain speech transcripts. The core of the system is formed by one-pass speech decoder performing time-synchronous Viterbi search. The acoustic model was discriminatively trained using the Minimum Phone

INTERSPEECH 2012 479
Error (MPE) criteria on a database containing 300 hours (from 3940 male and 2030 female speakers) of broadcast and close-talk microphone recordings. The model consists of tied-state context dependent HMMs of Czech phonemes and several types of non-speech events. It contains 4k of physical states with up to 32 components per state so that the total number of Gaussian components is approximately 120k.

The lexicon of the system contains 315k items, many of which have multiple pronunciations. The language model is based on bigrams estimated on a 10 GB corpus compiled from Czech (mainly newspaper) texts. We apply the modified Kneser-Ney discounting method for smoothing of the model.

3.1. Speaker adaptation

We use the Constrained Maximum Likelihood Linear Regression (CMLLR) for speaker adaptation. The method computes a feature transformation that aims to reduce the mismatch between an initial acoustic model and the adaptation data. Estimation of adaptation parameters consists of an iterative process where the likelihood of the adaptation data is being maximized with respect to a given preliminary transcription (which is the result of the first recognition pass). We apply only global CMLLR transformation for all Gaussians of the acoustic model. The non-speech events, as determined by the first recognition pass (taking into account values \( \alpha_e \)), are left out from the adaptation data. During the second recognition pass, adapted speaker-specific models are applied according to the smoothed diarization output. Segments attributed to any speaker are recognized with the initial speaker independent model.

4. Experiments and results

4.1. Datasets

Experiments were carried out using two test sets. The first set was created based on the COST278 multilingual pan-European broadcast news database. This set was used in order to demonstrate the effect of utilization of speech transcripts within the speaker diarization process even in the case of language mismatch of the ASR system. The second set was composed of recordings of Czech broadcast programs only. Training and development of the diarization system was common for both test sets and it was done using the held out part of the COST278 database. Utilization of the training data and development data followed the setup described in [7].

The COST278 test data comprised of 5 languages: Belgian Dutch, Czech, Hungarian, Slovak and Slovenian. The set consisted of 15 TV news shows of length in the range from 4 to 53 minutes (6.3 hours in total). The number of speakers in these shows varies in wide range from 2 up to 78.

The Czech test set consisted of 16 shows of length in the range from 4 to 51 minutes (4.3 hours in total). These recordings encompass both TV and radio programs. The number of speakers in these shows varies in range from 2 up to 55. In contrast to the first test set, the second set contains not just news but also interviews and thus more recordings contain less than 5 speakers in this case.

4.2. Evaluation metrics

Performance of diarization systems is usually evaluated by the Diarization Error Rate (DER). The DER can be decomposed as\( \text{DER} = \text{SPKE} + \text{FA} + \text{MISS} \), where the \( \text{SPKE} \), \( \text{FA} \) and \( \text{MISS} \) represent the speaker, speech false alarm and missed speech error rates respectively. A forgiveness region of 0.25 s (both + and -) was not scored around each boundary.

We also evaluate our system in terms of standard measures used to assess the segmentation performance. For their assessment, change-points detected by the system must be coupled with the reference ones first. A couple is constituted iff the detected change-point is the closest to the reference one and vice versa and, in addition, if the distance between them is smaller than 1 second. Then the recall, the precision and the F-rate measures are calculated respectively as

\[
R = \frac{H}{H + D}, \quad P = \frac{H}{H + I} \quad \text{and} \quad F = \frac{2RP}{R + P}
\]

where \( H \), \( I \), \( D \) denote the number of coupled, inserted and deleted change-points respectively.

Additionally, in order to highlight the effect of utilization of speech transcripts to improve the accuracy of positions of detected speaker turns, we assess the ratio of change-points that were detected within the intervals corresponding to the words uttered by speakers. We denote this ratio as the word-breakage (WB) rate. A forgiveness region of only 20 ms around the word-boundaries was used in this case.

Finally, speech recognition performance is evaluated in terms of the word-error-rate (WER).

4.3. Diarization system setup

The diarization system operates with feature vectors formed by 12 Mel-Frequency Cepstral Coefficients (MFCCs). The universal background GMM with 256 Gaussians was employed for extraction of sufficient statistics. We used 400-dimensional i-vectors and the LDA dimensional reduction to 200.

The UBM was trained using the data from 1007 speakers (2530 segments, 11.5 hours). The total variability space was estimated using a subset of the UBM training data resulting from the condition of minimal length of a segment of 3 seconds. This resulted in 2050 segments (10.2 hours) from 909 speakers. The LDA projection matrix was estimated using the data from speakers for which at least three segments of minimal length of 3 seconds are available, in total 1528 segments (7.5 hours) from 280 speakers were used.

4.4. Results

Tabs. 1 and 2 present achieved results in terms of both speaker segmentation and diarization performance. The system that makes no use of transcripts employs a SAD module that combines an energy and model (GMM) based detection. We point out that the segmentation measures were evaluated at the end of the diarization process (after the clustering stage). We conclude that utilization of speech transcripts in the diarization process yields improvement of both segmentation and clustering performance measures as determined by the higher F-rate and lower DER respectively. We can also conclude that incorporation of information about word boundaries yields remarkable reduction of the WB rate. Moreover, reduction of the number of change-point candidates leads to lower computational cost of the segmentation process as recognized by the real-time (RT) factor in the Tab. 1. In our experiments, the knowledge of oracle transcripts yields further slight improvement of the segmentation.

---

\(^2\)Please note that as the application of speaker adaptive recognition requires two speech recognition passes, the first recognition pass preceding the diarization process does not yield any extra burden.

\(^3\)Measured on a machine with Intel Core i7@2.66GHz.
Table 1: Segmentation and diarization performance for the multilingual COST278 test set. The WB rate was evaluated only for the Czech part of the set.

<table>
<thead>
<tr>
<th>Transcripts</th>
<th>R [%]</th>
<th>P [%]</th>
<th>F [%]</th>
<th>WB [%]</th>
<th>× RT</th>
<th>MISS [%]</th>
<th>FA [%]</th>
<th>SPKE [%]</th>
<th>DER [%]</th>
<th>× RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>87.5</td>
<td>53.8</td>
<td>66.6</td>
<td>49.9</td>
<td>0.14</td>
<td>1.8</td>
<td>0.6</td>
<td>11.5</td>
<td>13.9</td>
<td>0.05</td>
</tr>
<tr>
<td>yes</td>
<td>80.1</td>
<td>74.6</td>
<td>77.2</td>
<td>6.5</td>
<td>0.02</td>
<td>2.4</td>
<td>0.7</td>
<td>8.4</td>
<td>11.5</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 2: Segmentation and diarization performance for the Czech test set.

<table>
<thead>
<tr>
<th>Transcripts</th>
<th>R [%]</th>
<th>P [%]</th>
<th>F [%]</th>
<th>WB [%]</th>
<th>× RT</th>
<th>MISS [%]</th>
<th>FA [%]</th>
<th>SPKE [%]</th>
<th>DER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>83.2</td>
<td>52.8</td>
<td>64.6</td>
<td>42.3</td>
<td></td>
<td>2.9</td>
<td>3.2</td>
<td>15.6</td>
<td>21.7</td>
</tr>
<tr>
<td>yes</td>
<td>78.4</td>
<td>73.1</td>
<td>75.7</td>
<td>10.5</td>
<td></td>
<td>3.3</td>
<td>3.0</td>
<td>13.2</td>
<td>19.5</td>
</tr>
<tr>
<td>oracle</td>
<td>80.2</td>
<td>76.0</td>
<td>78.0</td>
<td>0.0</td>
<td></td>
<td>4.0</td>
<td>2.7</td>
<td>13.3</td>
<td>20.0</td>
</tr>
</tbody>
</table>

Table 3: Results of the SA system for the Czech part of the COST278 test set. The SI system achieved WER of 14.46 %

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Clustering</th>
<th>WER [%]</th>
<th>rel. impr. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed-blocks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>-</td>
<td>14.27</td>
<td>1.3</td>
</tr>
<tr>
<td>ASR-based</td>
<td>-</td>
<td>13.82</td>
<td>4.4</td>
</tr>
<tr>
<td>oracle</td>
<td>-</td>
<td>13.73</td>
<td>5.0</td>
</tr>
<tr>
<td>baseline</td>
<td>baseline</td>
<td>13.26</td>
<td>8.3</td>
</tr>
<tr>
<td>ASR-based</td>
<td>ASR-based</td>
<td>13.49</td>
<td>6.7</td>
</tr>
<tr>
<td>oracle</td>
<td>ASR-based</td>
<td>13.12</td>
<td>9.3</td>
</tr>
</tbody>
</table>

Table 4: Results of the SA system for the Czech test set. The SI system achieved WER of 24.41 %

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Clustering</th>
<th>WER [%]</th>
<th>rel. impr. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixed-blocks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>-</td>
<td>23.32</td>
<td>4.5</td>
</tr>
<tr>
<td>ASR-based</td>
<td>-</td>
<td>23.12</td>
<td>5.3</td>
</tr>
<tr>
<td>oracle</td>
<td>-</td>
<td>22.69</td>
<td>7.0</td>
</tr>
<tr>
<td>baseline</td>
<td>baseline</td>
<td>22.44</td>
<td>8.1</td>
</tr>
<tr>
<td>ASR-based</td>
<td>ASR-based</td>
<td>22.07</td>
<td>9.6</td>
</tr>
<tr>
<td>oracle</td>
<td>ASR-based</td>
<td>22.08</td>
<td>9.5</td>
</tr>
</tbody>
</table>

As well as adaptation for speaker homogeneous segments as identified by the speaker change-point detection. If the speech recognition accuracy is of the only interest, it is not important whether the automatic transcripts are taken into account during the diarization process or not. However, the main reason for utilization of transcripts is improvement of the quality of the diarization output.

6. Acknowledgments

This research work was supported by Czech Ministry of Culture (project no. DF11P01OVV013 in program NAKI).

7. References