The Use of Speaker Diarization Methods to Automatically Measure Gender Representation on Radio Podcasts

Seán Dexter

A dissertation presented for the degree of
Magister in Arte Ingenieria (MAI)

Department of Electrical and Electronic Engineering
Trinity College Dublin
Ireland
18th April 2017

The Use of Speaker Diarization Methods to Automatically Measure Gender Representation on Radio Podcasts

Seán Dexter

Abstract

Abstract goes here...

Contents

Chapter 1

Introduction

Introduction goes here...

1.1 Background

Background goes here...

1.2 Outline

Outline goes here...

Chapter 2

Literature Review

In this chapter, the present state of the art of diarization methods are discussed, highlighting areas pertinent to the primary purposes of this project. Other problems involved in broadcast news radio diarization, such as automatic removal of non-news segments, are also discussed.

2.1 Speaker Diarization

In 2006, Sue Tranter and Doug Reynolds provided an overview of automatic speaker diarization systems. They defined diarization as “the task of marking and categorising the audio sources within a spoken document” [1]. In their paper, they go on to outline the techniques commonly used for speaker diarization.

At the simplest level, diarization is described as the action of defining speech versus non-speech e.g. music, silence noise etc. Speaker diarization can be seen as a specification of this, where speaker changes are marked, and segments of speech (i.e. start and end times for parts of a speech document that mark transitions to or from non-speech segments, or between speaker segments) coming from the same speaker are labelled accordingly.

The three primary domains for speaker diarization research, defined by Tranter and Reynolds, are broadcast news audio, recorded meetings and telephone conversations. In the NIST Rich Transcription speaker evaluations, speaker diarization systems have focused on broadcast news and meeting data. Speaker diarization systems are also designed taking into account the amount of prior information needed for training.

In the case of broadcast news diarization, it is desirable for the system to have no prior information about the identity of speakers, and be able to perform diarization based on statistical reasoning methods applied to singular shows. Such systems are described by NIST [2], under the Rich Transcription criteria for evaluation.

2.2 Diarization Systems Evaluation

Since 2003, the NIST Rich Transcription Evaluation [2] has outlined the criteria for measuring error in speaker diarization systems each year, with little change from 2006 onwards. The evaluation documents outlined that there must be a one-to-one mapping between system output speakers and reference speakers, and that each file (or podcast in our case) was to be considered separately.

2.2.1 Diarization Error Calculation

The error in speaker diarization, upon which to base all diarization systems, is defined as the fraction of speaker time that is not attributed correctly to a speaker. This error rate is then broken down into three major categories:

Speaker Error Time

Any time that is attributed to the wrong speaker is counted as speaker error time. This is calculated by the following summation:

$$\frac{\sum_{all\ segs} dur(seg) * \{\ min(\ N_{Ref}(seg),\ N_{Sys}(seg)\ ) - N_{Correct}(seg)\ )\ \}}{\sum_{all\ segs} dur(seg) * N_{Ref}(seg)}$$

Where $dur(seg)$ is the duration of a segment, $N_{Ref}(seg)$ the number of reference speakers in the segment, $N_{Sys}(seg)$ the number of system speakers in the segment, and $N_{Correct}(seg)$ the number of speakers that are correctly matched between reference and system speakers. The other two forms of error can be described similarly, replacing the numerator summation term accordingly.

Missed Speaker Time

Missed speaker time is the sum of the following, over all segments where more reference speakers are speaking than system speakers:

$$dur(seg) * (\ N_{Ref}(seg) - N_{Sys}(seg)\ )$$

False Alarm Time

False alarm time is represented by the negative of the same sum as the missed speaker time, but over all segments where more system speakers are speaking than reference speakers. The sum of these three errors give the total diarization error.

2.2.2 Assessment of Systems Using Diarization Error

The IberSPEECH 2016 conference [3] and the ESTER-2 evaluation campaign [4] have based their standards of evaluation on these criteria provided by NIST, in order to judge the performance of systems participating as submissions. For IberSPEECH 2016, NIST's standard Perl script was required [2], for calculating the diarization error rate (DER), given that labels were in a specific format [3]. The Rich Transcription Time Marked format (RTTM) was described as having the following layout for each segment of labelled speech, to be passed as an input to the *md-eval-v21.pl* script.

SPEAKER	File	Ch	Beg_Time	Dur	<NA>	<NA>	Spkr_Label	<NA>	<NA>

Where:

1. SPEAKER = A tag indicating that the segment contains information about a single segment of speech belonging to one individual.
2. File = The name of the considered file.
3. Ch = Refers to the channel, which is defaulted to 1.
4. Beg_Time = The beginning time of the segment, in seconds, measured from the start of the file.
5. Dur = The duration of the segment, in seconds.
6. Spkr_Label = The label assigned to the speaker present in the considered segment.

Note here that the "<NA>" tags simply represent fields that are unused. Any time value in seconds was required to have a '.' decimal delimiter.

For IberSPEECH 2016, the criteria from NIST's 2006 standards of DER were considered [2], as opposed to the more recent criteria provided by NIST in 2009 [5]. What this meant for the DER calculation was that 250 millisecond time collars were employed around each reference segment, in order to forgive timing errors in the reference. This change, unique to the 2006 definitions, was due to segment times in the reference data provided at the time that were not created via a high accuracy process [2], and therefore would have required for leniency in the error calculation.

The Albayzin IberSPEECH 2016 guidelines also specify precisely how the command for how the *md-eval-v21.pl* Perl script should be run in bash:

```
perl md-eval-v21.pl -c 0.25 -r <REFERENCE>.rttm -s <SYSTEM_OUTPUT>.rttm
```

It is important to note here that there is a slight difference between NIST's and IberSPEECH 2016's guidelines here i.e. some tags are left out. This is likely due to the version of the Perl script mentioned being older than what is available now.

2.3 Examples of Diarization Systems

For broadcast news diarization, there exist many systems that have been built to achieve low error rates

2.4 DCU

This is advantageous, since the reference data from [6] is only accurate to the nearest whole second.

[6] The reference data from DCU Political Science students...

Chapter 3

Diarization and Gender Detection

3.1 The LIUM SpkDiarization Toolkit

In order to perform Diarization on any test podcast, the LIUM SpkDiarization toolkit needed to be run in such a way as to perform Single-show Broadcast News Diarization. A bash shell script, provided on the LIUM website [7], listed the necessary Java commands in order to perform Diarization on one broadcast news podcast, with no training data. The only input to this script is the podcast .WAV file. The commands were as follows:

- Speech/Music/Silence Segmentation (short-named PMS)
- GLR Based Segmentation
- Linear Clustering
- Hierarchical Clustering
- Initialize GMM
- EM computation
- Viterbi Decoding
- Adjust segment boundaries
- Filter Spk Segmentation According to PMS Segmentation
- Split segment longer than 20s
- Set gender and bandwith
- CLR clustering

The result of all this, is a .seg file. Each line in this file corresponds to a segment of speech. For example:

show	1	0	480	F	S	U	S4

Where:

1. show = The name of the input file, without the extension.
2. 1 = The channel number. This was always kept at 1.
3. 0 = The start of the segment in centiseconds.
4. 480 = The duration of the segment in centiseconds.
5. F = The identified gender of the speaker; F, M or U (unknown)
6. S = The type of band; S for studio, T for telephone.
7. U = The type of environment i.e. music, speech only, etc.
8. S4 = The speaker label.

Comments were denoted with two semi-colons (;;). This became important to note when rearranging the format of the .seg files.

3.1.1 Evaluation Methods

The Diarization Error Rate, as described in (SECTION HERE), is the standard that is used to measure accuracy of a Speaker Diarization system such as LIUM. The segmentation scoring tool described requires the following for each segment, in a Rich Transcription Time Marked (.RTTM) file:

SPEAKER	File	Ch	Beg_Time	Dur	<NA> <NA>	Spkr_Label	<NA> <NA>

Where:

1. SPEAKER = A tag indicating that the segment contains information about a single segment of speech belonging to one individual.
2. File = The name of the considered file.
3. Ch = Refers to the channel (kept at 1, as seen previously for LIUM).
4. Beg_Time = The beginning time of the segment, in seconds, measured from the start of the file.
5. Dur = The duration of the segment, in seconds.
6. Spkr_Label = The label assigned to the speaker present in the considered segment.

Note here that the "<NA>" tags simply represent fields that are unused. Any time value in seconds must have a '.' decimal delimiter.

The freely available "md-eval-vXX.pl" Perl script (REFERENCE), where "XX" represents the version number, could be executed with the following bash shell command:

```
perl md-eval-v21.pl -c 0.25 -r <REFERENCE>.rttm -s <SYSTEM_OUTPUT>.rttm
```

This would then compute the Diarization Error Rate and print out all pertinent information about each component of the DER.

3.2 Testing LIUM on Fixed Datasets

The first set of data that LIUM was tested on consisted of fifty segments of speech from well known Irish radio presenters, lasting 30 seconds each. Half of these presenters were male, and the other half, female. A 25 minute podcast made up entirely out of this data was then used to test LIUM. A reference file was created following the aforementioned .RTTM format, labelling speaker time for every 30-second segment.

The output .seg file from LIUM was converted into .RTTM format by appending the AWK command shown below to the bash shell script for single-show Diarization. Firstly, the comment delimiter (;;) was defined, and then each variable could be converted accordingly. The “$8” here represents the speaker label, which was replaced by a “$5” when only considering the error in gender detection. The “$show” represented the base-name of the .wav file that the LIUM Diarization script was run on.

```
awk '!/^;;/ {print "SPEAKER " $1 " " $2 " " ($3/ 100) " " ($4/ 100) " <NA> <NA> " $8 "
    <NA> <NA>"}' $show.c.3.seg > $show.rttm
```

The Albayzin Perl script could then be run on both .RTTM files, and the Diarization Error Rate calculated for speaker and gender-only scenarios.

3.2.1 Varying Age

The original data contained variation in the age of the speakers. LIUM was run on this test podcast, and the output was converted into .RTTM format. This small test showed that age difference had an impact in how a statistical speaker Diarization system would find it difficult to identify a speaker as the same from two segments of speech from when their age was different by about 10 years.

3.2.2 Static Age

In order to eliminate the varying factor of age in the speakers, five minutes of older speech segments, from when the presenters were younger, were removed from the made-up podcast data. LIUM was then run again on the new podcast without any variance in age.

3.3 Introducing the DCU Reference Data

As seen in (SECTION), the data from DCU was simplified so that each line contained the following:

show	week	day	radio	name	gender	start	end	dur

The following AWK command could then be run to convert this from its raw .csv format into .RTTM, with an extra ".ref" tag to denote that this is the reference file for the Albayzin Perl script. As before, the "$5" was replaced with "$6" in order to measure error in gender detection.

```
awk {print "SPEAKER " $show " 1 " $7 ".00 " $9 ".00 <NA> <NA> " $5 " <NA> <NA>"}'
    $show.csv > $show.ref.rttm
```

3.4 Fixing the Reference Data

3.4.1 Adverts

3.4.2 No Adverts

3.5 Exploitation of DCU Data to Remove Adverts

3.6 Automatic Non-news Detection Algorithm

3.6.1 Root Mean Square Distance

RMS distance

3.6.2 Arithmetic-Harmonic Sphericity Distance

AHS distance

Chapter 4

The Automatic Detection of Non-News Segments

4.1 All References (CHANGE)

[6] [4] [8] [9] [3] [10] [11] [12] [13] [14] [2] [5]

Chapter 5

Conclusion

My Conclusion goes here

Appendix A

Appendix Title One

Bibliography

[1] Sue E Tranter and Douglas A Reynolds. “An overview of automatic speaker diarization systems”. In: *IEEE Transactions on audio, speech, and language processing* 14.5 (2006), pp. 1557–1565.

[2] NIST. *Rich Transcription Spring 2006 Evaluation*. URL: http://itl.nist.gov/iad/mig/tests/rt/2006-spring/index.html.

[3] Alfonso Ortega et al. *The Albayzin 2016 Speaker Diarization Evaluation*. 2016. URL: https://iberspeech2016.inesc-id.pt/index.php/albayzin-evaluation/.

[4] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. “The ester 2 evaluation campaign for the rich transcription of French radio broadcasts.” In: 2009.

[5] NIST. *Rich Transcription 2009 Evaluation*. URL: http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html.

[6] Kathy Walsh et al. “Hearing Womens Voices?: Exploring womens underrepresentation in current affairs radio programming at peak listening times in Ireland”. PhD thesis. DCU, 2015.

[7] Sylvain Meignier et al. *LIUM Speaker Diarizaition Wiki*. 2013. URL: http://www-lium.univ-lemans.fr/diarization/doku.php/.

[8] Mickael Rouvier et al. *An open-source state-of-the-art toolbox for broadcast news diarization*. Tech. rep. Idiap, 2013.

[9] Sylvain Meignier and Teva Merlin. “LIUM SpkDiarization: an open source toolkit for diarization”. In: *CMU SPUD Workshop*. Vol. 2010. 2010.

[10] S. E. Tranter and Douglas A. Reynolds. “Speaker diarisation for broadcast news”. In: *ODYS-2004, 337-344.* (2004).

[11] Sue E Johnson et al. “Spoken Document Retrieval for TREC-8 at Cambridge University.” In: *TREC*. Vol. 2. 3.1. 1999, pp. 3–1.

[12] Frederic Bimbot and Luc Mathan. “Text-free speaker recognition using an arithmetic-harmonic sphericity measure”. In: *Third European Conference on Speech Communication and Technology*. 1993.

[13] David Tavárez et al. “Aholab Speaker Diarization System for Albayzin 2016 Evaluation Campaign”. In: *IberSPEECH 2016: Lisbon, Portugal Proceedings*. 2016.

[14] Jose Patino et al. “EURECOM submission to the Albayzin 2016 Speaker Diarization Evaluation”. In: *IberSPEECH 2016: Lisbon, Portugal Proceedings*. 2016.