NUTS AND FLAKES: A STUDY OF DATA CHARACTERISTICS IN SPEAKER DIARIZATION

Nikki Mirghafori and Chuck Wooters

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704
{nikki,wooters}@icsi.berkeley.edu

ABSTRACT

Researchers in the speaker diarization community have observed that some audio files show unusually high Diarization Error Rates (DER) (hard to crack *"nuts"*), and some exhibit hyper-sensitivity to tuning parameters (*"flakes"*). The goal of this study is to systematically study the features that correlate with such behavior. We calculated over forty features for each of 24 shows from the Broadcast News corpus along the dimensions of speaker count, conversation turn, and speaker and show duration. We observed that number of speakers, number of turns, and do-nothing DER (a measure related to the percentage of time the dominant speaker spoke) correlated best with "nuttiness". The do-nothing DER and number of speakers were also the best correlates of "flakiness".

1. INTRODUCTION

The goal of speaker diarization is to segment an audio recording into speaker-homogeneous regions [11]. Typically, this segmentation must be performed with little knowledge of the characteristics of the recording or of the talkers in the recording. For example, we may know the source and date of the audio recording (e.g. CNN Nightly News), but we typically do not know how many speakers occur in the recording, whether one speaker is speaking most of the time, how many males vs. females, whether there is music in the recording, etc.

In recent years, NIST has held evaluations of speaker diarization technology [5, 6]. The measure of performance that NIST uses for diarization systems is *Diarization Error Rate* (DER). DER is calculated by first finding the optimal match between the true speakers and the hypothesized speakers and then calculating the percentage of time that is incorrectly assigned according to the optimal match (see Figure 1.)

Because DER is a time-based metric, the error will tend to be dominated by the speakers who speak the most. Thus, for diarization systems that use agglomerative clustering (as most do), achieving the optimal DER critically depends on making the right decision about when to stop clustering. For shows that are dominated by one (or a few) speakers, stopping too early or too late can lead to very high DERs. In addition to DER, other measures of speaker diarization performance have been used. For example, in [1] they proposed a measure that is a combination of average speaker purity and average cluster purity. In the Fall 2003 NIST evaluations [3], NIST used a speaker diarization measure that was based on the number of words that were correctly assigned to each speaker (known as the "Who Spoke The Words" evaluation.) Despite the inherently noisy nature of DER, it has the advantage that it does not make any assumptions about the possible down-stream applications of speaker diarization and has become the standard measure of speaker diarization performance.

This work was partly supported by the Defense Advanced Research Projects Agency under Contract No. MDA972-02-C-0038 as part of a subcontract to ICSI by SRI International. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

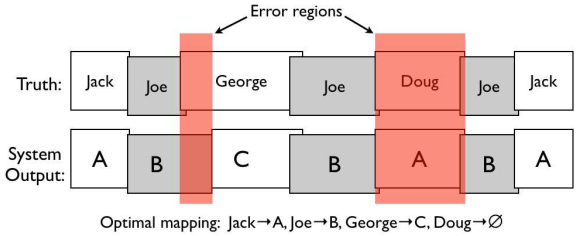

Fig. 1. Scoring speaker diarization systems.

At the EARS Fall 2004 Workshop (RT04) held in Palisades NY [4], several researchers working on speaker diarization reported that some of the Broadcast News (BN) shows exhibited DER hyper-sensitivity. That is, changes in the parameter settings of a diarization system would result in dramatic swings in the DER for some shows. For example, Figure 2 gives the DERs for two different BN shows. We ran each show with eight different parameter settings of our speaker diarization system. The first show (CNBC) has relatively low variation across the eight runs, while the second show (VOA) demonstrates parameter hyper-sensitivity.

This effect was also reported in [13] where changing a system parameter resulted in a lower DER for five out of six shows. The error rate on the sixth show almost tripled resulting in a higher overall DER making the system appear worse than it really was.

At the RT04 Fall workshop it was also noted that some shows were much more difficult to diarize than other shows. There have been various studies to analyze the correlates of error rate for Automatic Speech Recognition (ASR) systems and these studies have suggested factors such as noise, fast speech [9], Lombard effect, the beginning and ending of sentences [2], etc. However, no such studies have been performed in the younger field of speaker diarization.

In this paper we examine characteristics of BN shows in order to answer two questions:

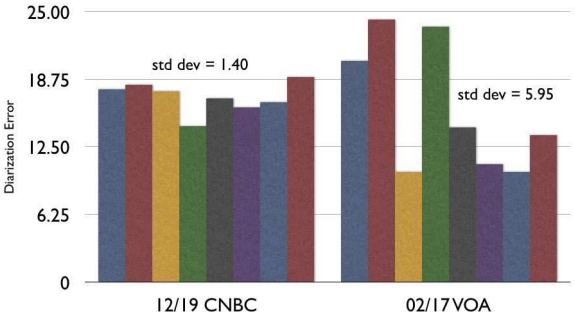

Fig. 2. Variation in DER for two BN shows. Each bar represents the DER for the show obtained by running the ICSI speaker diarization system with a particular choice of parameter settings. The parameters that we varied were: number of initial clusters, number of initial gaussians per cluster, and minimun duration of a cluster.

- Which show characteristics are associated with higher DER (*hard to crack "nuts"*)?
- Which show characteristics are associated with parameter hyper-sensitivity (*"flakes"*)?

To perform this analysis, we have collected speaker diarization scores[1] from five research groups: ICSI [14], MIT-LL [10], Cambridge University [12], LIMSI [15], and LIA [7, 8]. We used one (representative) set of DER scores provided by each site to examine the issue of show difficulty. To examine the issue of parameter hyper-sensitivity, we used several sets of scores, generated from multiple runs of a site's diarization system.

In Section 2 we discuss the BN data used in this study and the data we received from each of the sites. In Section 3, we discuss the features we extracted from each of the BN shows for performing the correlation analysis. In Section 4 we present our analysis of the data and in Section 5 we present our conclusions.

2. DATA

2.1. Show Data

The data used in this study consist of roughly 30 minute excerpts from 24 BN shows (12 Dev04F and 12 Eval04F.) All shows were recorded from a variety of broadcast sources including PRI, NBC, CNN, CSPAN, ABC, PBS, etc. The Eval04F shows were recorded during December 2003 and the Dev04F shows were recorded during February 2001 and November and December 2003. Much of the data consists of news anchors reading the news, but it also includes background sounds, music and other talkers (reporters and interviewees.)

2.2. System Data

We asked each of the sites to provide us with data from their speaker diarization systems. The data is split into two basic types:

1. One DER score for each of the 24 shows, produced from the best (or a representative) configuration of their system. Ideally, the particular system configuration would be chosen based on the overall DER on the *Dev04F* shows.
2. Multiple sets of DER scores for each of the 24 shows. These would ideally be produced by running a system with different parameter settings. For example, the ICSI data were generated by varying the following parameters: number of initial clusters, number of initial gaussians per cluster, and minimum duration of a cluster.

The type 1 data were used to study show difficulty and the type 2 data were used to study show flakiness. Not all sites could provide all of the requested data. For the type 1 data, three sites were able to provide scores for all 24 shows and two sites provided data for 12 shows. For the type 2 data, we were able to use scores from multiple system runs from two sites.

3. FEATURES

Our goal was to characterize a show by calculating relevant features, attaining maximal coverage as well as parsimony. We calculated many features, but after observation, eliminated some of the confounded ones. For the sake of completeness, we report on all the features we considered.

We calculated features pertaining to *speaker count, conversation turns, and speaker and show duration*, as follows:

Speaker count features: For each show, we calculated the total number of speakers, number of male speakers, number of female speakers, ratio of male speakers, and ratio of female speakers. These speaker count features were included as there had been anecdotal observations that shows with higher number of speakers incur a higher DER. The number and ratio of females and male speaker features attempted to get at gender effects, if any.

Conversation turn features: We considered the number of conversation turn changes per minute, total number of turns in the show, mean and standard deviation of turn durations, and the normalized entropy of turn durations. Each turn feature was calculated with four different window lengths of 0.5, 1, 2, and 3 seconds. For example, for the window of 2s, if the speaker stopped talking (to catch her breath) or was interrupted (by back-channel or laughter) for less than 2s, the pre- and post-interruption segments would be considered as part of the same turn. However, if the interruption was longer than 2s, they would be considered to be different turns. No distinctions for type of interruptions were made.

In the analysis stage, we observed that the features calculated using a window length of 0.5s correlated the least and the those for 2s and 3s were usually the best. Given the high level of correlation between these features, only the 2s window-size features were retained.

Speaker duration features: We calculated the normalized entropy of total speaker duration, the "do-nothing" score, and the percentage of scored show-time that the dominant N (where, N ranged from 1 to 3) speakers spoke.

A small value for normalized entropy of total speaker duration indicates that there were a few dominant speakers in the show (e.g., show anchors) and the rest of the speakers spoke relatively little. Entropy was normalized (divided) by the maximum possible entropy, since the maximum possible entropy (where all speakers speak equal amounts) is larger if there are more speakers in a show. The un-normalized entropy correlated highly ($\rho = 0.87$) with the number of speakers and, when used in linear regression with

[1] As we are not interested in making comparisons between the different sites, we have anonymized all of the scores presented in this paper.

cross product terms, was a less predictive feature than the combination of the normalized entropy *and* the number of speakers. It was thus eliminated.

The "do-nothing" score is the DER of a show if all the data are assigned to the dominant cluster, which in our case, was calculated using the truth files. This score is "synoymous" (ρ=-1, when one is large, the other is small) with the percentage of time spoken by the top dominant speaker. The proportion of time spoken by the top two and three speakers proved to be a confounded measure, as, for example, a show with three speakers will have a feature value of 100% for the percent time top three speakers spoke, and essentially encode the total number of speakers (ρ=-0.78).

Show duration features: Total show duration, duration of scored regions, duration of non-scored regions, and ratio of duration of non-scored regions were calculated. Some segments of the shows (e.g., commercials and music) are run through the diarization system, but are not scored. Features pertaining to the duration of non-scored regions were aimed to address the potential error and cluster impurity that this may have caused.

4. ANALYSIS

4.1. Correlates of "Nuttiness"

As we see in Figure 3, there seems to be general agreement between sites as to which shows are harder to diarize. For three sites, we had the DER for all 24 shows (both Dev04F and Eval04F) and for two sites, we had DER scores for only the 12 Eval04F shows.

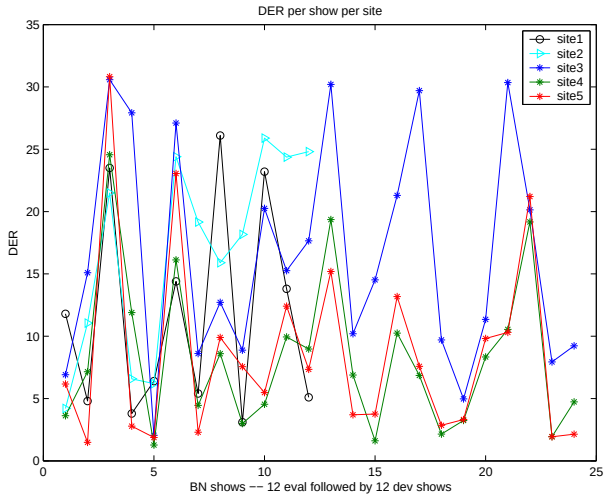

Fig. 3. *BN Eval04F and Dev04F DER for all systems.*

In order to find correlates of "nuttiness" (i.e., exhibition of high DER) without bias toward the internals of any site-specific system, we pooled the representative DERs from all five systems and calculated the mean DER for each show across sites. As mentioned in Section 1, DER tends to be an inherently noisy measure, so calculating the mean over all sites provides some desirable smoothing. Yet, with such limited number of data points (24), the observations should mainly be considered as suggested data patterns.

We calculated Spearman rank correlation coefficients between all the features and the mean DER. The advantage of the Spearman correlation coefficient is that it captures non-linear correlations between the data, especially given the noisy nature of DER. Table 1 shows the correlation coefficients (ρ) and the p-values (the smaller the p-value, the more significant the correlation), for the show features and the all-site mean DER. The top three "independent" factors which correlate the most with DER are: the total number of speakers, the total number of turns, and the do-nothing DER. The number of females and males are tied to the total number of speakers. Given that the duration of the BN shows are similar, the number of turns, turn duration mean, and the number of turns per minute are related measures. So are the do-nothing score and the percentage of time the dominant speaker spoke.

Show Feature	**Spearman ρ (p-value)**
Number of speakers	+0.77 (0.00001)
Number of females	+0.72 (0.00008)
Number of males	+0.60 (0.00200)
Number of turns	+0.58 (0.00289)
Turn duration mean	-0.57 (0.00380)
Do-nothing DER	+0.56 (0.00436)
Turns per minute	+0.55 (0.00577)
Dominant speaker duration %	-0.52 (0.00976)
Ratio of males	-0.41 (0.04927)
Ratio of females	+0.39 (0.06178)
Non-scored duration	+0.34 (0.10032)
Turn duration std. dev.	-0.32 (0.12632)
Ratio of non-scored	+0.30 (0.14941)
Scored duration	-0.23 (0.28810)
Normalized turn entropy	+0.18 (0.40930)
Normalized speaker entropy	-0.05 (0.83382)
Show duration	-0.02 (0.92289)

Table 1. The Spearman rank correlations between the all-site mean DER and the BN show features.

We note that DER mean and standard deviation had a correlation of ρ=+0.77 (p-value = 0.0001) with one another: the higher the DER, the more the show will exhibit variability from site to site.

We also ran regression on individual show features and compared the mean squared errors (MSE). The MSE rankings suggested a similar (almost identical) order of importance to the features. We attempted to run regression using multiple features, however, given the limited number of data points, the calculations ran into matrix singularity problems. We also constructed regression trees using only the top three independent features. The root node, as expected, was the number of speakers. The nodes in the lower regions of the tree, however, seemed noisy due to data sparsity.

4.2. Finding "Flakiness"

As mentioned in Section 2, we had DERs from multiple runs (13 of ours and 12 of another site's), each with a different parameter setting. A show is considered "flaky" if it has a large DER standard deviation. To calculate the DER standard deviation, we subtracted the show-and-system-specific mean from each show's DER and pooled the data from the two sites. This normalization was necessary so that variation due to the difference in the range of scores across systems did not artificially inflate the observed "flakiness".

Show Feature	Spearman ρ (p-value)
Do-nothing DER	+0.51 (0.01066)
Dominant speaker duration %	-0.49 (0.01436)
Number of speakers	+0.47 (0.02176)
Number of females	+0.43 (0.03842)
Number of males	+0.38 (0.07025)
Number of turns	+0.28 (0.17949)
Ratio of males	-0.25 (0.23109)
Ratio of females	+0.18 (0.40242)
Turn duration mean	-0.15 (0.49319)
Show duration	+0.15 (0.49578)
Turns per minute	+0.13 (0.54896)
Non-scored duration	+0.10 (0.65075)
Normalized speaker entropy	+0.09 (0.68620)
Scored duration	+0.08 (0.71020)
Normalized turn entropy	+0.06 (0.76515)
Ratio of non-scored	+0.06 (0.76515)
Turn duration std. dev.	+0.04 (0.83698)

Table 2. The Spearman rank correlations between the multi-site DER standard deviation and the BN show features.

Table 2 shows the Spearman correlation coefficients (ρ) and the p-values for the show features and the multi-site DER standard deviation. The correlations are neither as strong nor as significant as in case of "nuts". The factor with the highest ρ is the do-nothing DER, which is followed closely by the related feature of percentage speech time of the dominant speaker. The next feature of relative significance is the total number of speakers. Speaker turn features do not appear to be as significant in determining "flakiness".

5. CONCLUSIONS

It appears that the "nuts" tend to have many speakers, a large number of speaker turn changes (and therefore, short turn durations and a high turns-per-minute rate) and a high do-nothing DER (i.e., the dominant speaker is not voluble). The correlation of these factors with high DER is compelling, as long uninterrupted speech segments spoken by only a few speakers seem intuitively easier to diarize than frequently interrupted short segments from many speakers where no one speaker is dominant.

The relationship between "flakiness" and do-nothing DER and number of speakers is weaker, but it may be that if there are a few speakers and the dominant speaker speaks much of the time, most diarization systems identify and cluster that speaker correctly and the DER is stable for various tuning parameters.

The current study has been done with very limited data, and more data points (BN shows) are needed to strengthen the correlation observations. Increased data will also allow the prediction of DER and its variability based on the show features through regression. This prediction, however, would be mainly of academic interest. The goal of this study has been to attempt to shed light on data characteristics, without focusing on any particular system, to help improve diarization accuracy. These observations have already suggested ideas to improve our diarization system, and we hope they prove to be helpful to others in eventually cracking the "nuts" and managing the "flakes".

Acknowledgments

This paper would not have been possible without data contribution from our colleagues: We would like to thank Corinne Fredouille (LIA) and Claude Barras (LIMSI) for sharing their data with us, and Sue Tranter (CUED) and Doug Reynolds (MITLL) for both sharing their data and, most especially, their insightful comments and suggestions. For the latter, we are also grateful to George Doddington. Finally, thanks to Steven Ketchpel for editorial comments.

6. REFERENCES

[1] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan. Unknown-multiple speaker clustering using HMM. In *ICSLP*, Denver, Colorado, USA, September 2002.

[2] N. Duta and R. Schwartz. Error analysis of the BN and CTS results. In *EARS STT workshop*, St. Thomas, U.S. Virgin Islands, December 2003.

[3] National Institute for Standards and Technology. NIST 2003 Fall Rich Transcription Evaluation website. http://www.nist.gov/speech/tests/rt/rt2003/fall.

[4] National Institute for Standards and Technology. NIST 2004 Fall Rich Transcription Evaluation website. http://www.nist.gov/speech/tests/rt/rt2004/fall.

[5] National Institute for Standards and Technology. NIST Rich Transcription evaluations. http://www.nist.gov/speech/tests/rt.

[6] National Institute for Standards and Technology. NIST Spring Rich Transcription Evaluation in Meetings website. http://www.nist.gov/speech/tests/rt/rt2005/spring.

[7] D. Istrate, N. Scheffler, C. Fredouille, and J.-F. Bonastre. Broadcast news speaker tracking for ESTER 2005 campaign. In *INTERSPEECH*, pages 2445–2448, Lisbon, Portugal, September 2005.

[8] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language (to appear)*, 2005.

[9] N. Mirghafori, E. Fosler, and N. Morgan. Why is ASR harder for fast speech and what can we do about it? In *ASRU*, pages 179–183, Snowbird, Utah, December 1995.

[10] D.A. Reynolds and P. Torres-Carrasquillo. The MIT Lincoln Laboratories RT-04F diarization systems: Applications to broadcast audio and telephone conversations. In *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY, November 2004.

[11] D.A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *ICASSP*, pages 953–956, Philadelphia, PA, March 2005.

[12] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland. The Cambridge University March 2005 speaker diarisation system. In *INTERSPEECH*, pages 2437–2440, Lisbon, Portugal, September 2005.

[13] S. E. Tranter. Two-way cluster voting to improve speaker diarisation performance. In *ICASSP*, pages 753–756, Philadelphia, PA, March 2005.

[14] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY, November 2004.

[15] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain. Combining speaker identification and BIC for speaker diarization. In *INTERSPEECH*, pages 2441–2444, Lisbon, Portugal, September 2005.