# Speaker Diarization Using Linguistic Information

*Leonardo Canseco-Rodriguez, Lori Lamel, Jean-Luc Gauvain*

Spoken Language Processing Group, (http://www.limsi.fr/tlp)
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France

{lcanseco,lamel,gauvain}@limsi.fr

## 1 Introduction

This communication explores the significance of linguistic information in speaker diarization from automatic broadcast news transcripts. The content of a broadcast news program is a rich source of information that in many cases reveals the true identity of those who take part in the show. It also includes information about the roles of the speakers by indicating who is the anchor and who are the reporters. Also, it provides information about the topic structure of the show given in the headlines and in announcements of commercial breaks, as well as specific formulations to signal the beginnings and ends of stories. The single or combined use of these three mains types of information allows a broadcast news audio recording to be structured into individual news stories for further diarization.

## 2 Linguistic-based Diarization

The typical broadcast news show has an anchor who leads the program, introducing reporters, guests and commercial breaks. Such speaker introductions occur frequently, revealing the true names of who speaks. Linguistic patterns were developed to detect the next, previous and current speaker. Rules for each pattern are used associate the true speaker names with the speech segments. In order to identify weakness and strengths of a linguistically-based approach, the diarization is applied to manual and automatic transcripts. In addition to comparing performances with perfect and imperfect transcripts, this comparison allows us to learn which linguistic information useful for diarization is missing in the automatic transcripts. After tagging entities to the transcripts, the most frequent patterns which provide information about the speaker are classified according to the situations where they appear. Such situations mainly correspond to announcements of who is speaking (self-speaker rules), who will speak (next speaker rules) or who just spoke (previous speaker rules). Details of the complete process can be found in [Can04].

## 3 Diarization Experiments

Since the effectiveness of linguistic patterns for diarization depends on the quality of the transcription, the performance using automatic transcripts generated with

**Table 1.** Diarization error rates using linguistic patterns on manual and automatic *Hub-4e* transcriptions.

| *Evaluation* | *Manual Transcription* | | | *Automatic Transcription* | | |
|---|---|---|---|---|---|---|
| *Cases* | *self-spkr* | *next-spk* | *prev-spkr* | *self-spkr* | *next-spk* | *prev-spkr* |
| *#corr1* | 2137 (98.5%) | 1186 (73.5%) | 135 (18.4%) | 1239 (75.2%) | 756 (64.2%) | 85 (20.2%) |
| *#corr2* | - | - | - | 75 (4.5%) | 73 (6.2%) | 6 (1.4%) |
| *#corr3* | 28 (1.2%) | 209 (12.9%) | 390 (53.2%) | 231 (14%) | 150 (12.7%) | 154 (36.6%) |
| *#corr4* | - | - | - | 18 (1%) | 10 (0.8%) | 10 (2.3%) |
| *#False id* | 4 (0.1%) | 217 (13.4%) | 208 (28.3%) | 84 (5.1%) | 188 (15.9%) | 165 (39.2%) |
| *#undef.* | 81 | 146 | 119 | 73 | 111 | 74 |
| *Tot.Matches* | 2250 | 1758 | 852 | 1976 | 1474 | 552 |

**Table 2.** Diarization error rates using linguistic patterns on manual and automatic transcripts (*97-98-99 Hub4-e* evaluation data).

| *Evaluation* | *Manual Transcriptions* | | | *Automatic Transcription* | | |
|---|---|---|---|---|---|---|
| *Cases* | *self-spkr* | *next-spk* | *prev-spkr* | *self-spkr* | *next-spk* | *prev-spkr* |
| *#corr1* | 115 (95%) | 50 (54.9%) | 7 (19.1%) | 94 (81%) | 38 (54.4%) | 8 (28.6%) |
| *#corr2* | - | - | - | 2 (1.7%) | 3 (8.6%) | - |
| *#corr3* | 7 (4.9%) | 22 (24.8%) | 18 (38.5%) | 7 (8.7%) | 10 (16%) | 11 (25.1%) |
| *#corr4* | - | - | - | - | - | - |
| *#False id* | - | 16 (20.2%) | 19 (42.2%) | 9 (8.4%) | 12 (20.8%) | 19 (46.1%) |
| *#undef.* | - | 3 | 1 | - | 2 | 1 |
| *Tot.Matches* | 122 | 91 | 45 | 112 | 65 | 39 |

an LVCSR system [Gau02] are compared with those obtained using manual transcriptions. The LVCSR acoustic models were trained on about 140 hours of data from the *TDT2* corpus using a lightly supervised approach, and the language model was estimated on about 1 billion of words of texts.

Table 1 summarizes the performance of the self-speaker, next-speaker and previous-speaker rules when these are applied to manual and automatic transcriptions of the *Hub-4e* corpus. The self-speaker rule largely outperforms the other rules having the lowest false identity association rate, and the previous-speaker rule has the highest one. The total number of identity associations done by the three rules, is reduced of about 18% for the automatic transcription when comparing with that one for the manual transcripts. And the total number of false identity associations done by the three rules, represents of about 9% of the total number of association done; this percentage is the same for both transcriptions. The same linguistic patterns and rules were tested on about 10 hours of unseen data on the NIST evaluation sets as shown in Table 2.

## References


[Can04] Canseco L., Lamel L., and Gauvain, J.L., "Speaker Diarization From Speech Transcripts", ICSLP'04.

[Gau02] Gauvain, J.L., Lamel, L., Adda, G., "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, 2002.