Speaker Diarization from Speech Transcripts

Leonardo Canseco-Rodriguez, Lori Lamel, Jean-Luc Gauvain
Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lcanseco,lamel,gauvain}@limsi.fr

ABSTRACT

The aim of this study is to investigate the use of the linguistic information present in the audio signal to structure broadcast news data, and in particular to associate speaker identities with audio segments. While speaker recognition has been an active area of research for many years, addressing the problem of identifying speakers in huge audio corpora is relatively recent and has been mainly concerned with speaker tracking. The speech transcriptions contain a wealth of linguistic information that is useful for speaker diarization. Patterns which can be used to identify the current, previous or next speaker have been developed based on the analysis of 150 hours of manually transcribed broadcast news data. Each pattern is associated with one or more rules. After validation on the training transcripts, these patterns and rules were tested on an independent data set containing transcripts of 10 hours of broadcasts.

1. INTRODUCTION

This paper describes recent studies on speaker diarization for broadcast news data. The goal is to use the linguistic information present in the speech transcripts to associate speaker identities with audio segments. The basic idea is that broadcast news shows can be structured into a series of segments (reports, interviews) which in turn can be characterized by their dynamics (narratives, speech extracts, interactions). Narratives generally correspond to segments spoken by the news anchor (or moderator) or on-site reporters. Speech extracts occur during press reports where a portion of audio data recorded at some prior event is played during the broadcast. Such extracts can be statements by well-known public figures (political speeches, prepared commentaries by civil authorities) or first-hand witnesses to some event. Interactions refer to dialogs concerning two or more speakers, where there is an exchange of ideas, often with explicit questions and answers. Extracts can be differentiated from interviews in that there is no exchange between the speaker and the moderator or reporter.

While speaker recognition has been an active area of research for many years, addressing the problem of identifying speakers in huge audio corpora is relatively recent. Most of the research has been concerned with speaker tracking (identifying and regrouping segments from the same speaker) within a single audio document, where the audio document can correspond to a radio or television broadcast [1, 3, 4, 5, 6, 7], a public hearing[6] or a telephone conversation [2, 7]. Speaker tracking is a part of what is often referred to as audio data partitioning, which aims to divide the acoustic signal into homogeneous, non-overlapping segments, identifying and removing non-speech segments, and associating document-relative speaker identities with each speech segment [3]. In NIST metadata evaluations, reported speaker tracking error rates are in the range of 10-20%.

Figure 1 shows examples of typical linguistic patterns which can be used to identify speakers in a news broadcast. Segment A corresponds to the anchor, who introduces herself by saying "hello, I am Joie Chan" and introduces the upcoming reporter (Segment B) with the sentence "Candy Crowley has the report". The reporter finishes her report and signs off with "for CNN this is Candy Crowley" and passes the control back to the anchor (Segment E) who thanks the reporter. This work explores the use of the lexical information, which can be combined with speaker tracking to assign the true speaker identities to speech segments.

2. SPEAKER NAME PATTERNS

The corpus of broadcast news shows used in this study was distributed by the Linguistic Data Consortium, and contains 150 hours of audio data from a variety television and radio sources, annotated with detailed manual transcriptions. The data were broadcast from 1993 through 1998, and come from a variety of sources: ABC (Nightline, World News Now, World News Tonight), CNN (Early Edition, Early Prime, Prime Time Live, Headline News, Prime News, The World Today), CSPAN (Washington Journal, Public Policy), and NPR (All Things Considered, Marketplace). The manual transcriptions specify the true speaker names when identifiable, and use distinct identifiers (spkr1, johndoc1, janedoe1) for speakers who are not known. True speaker identities are known for about 40% of the distinct speakers, accounting for almost 85% of the data. These data were used to identify linguistic patterns for speaker names and to validate the rules. The 1997, 1998 and 1999 DARPA/NIST evaluation test sets are then used to assess the approach on about 10 hours of unseen data from the same epoch.

From the example in Figure 1 it is clear that there are frequently occurring linguistic patterns that can be useful for identify the current, the previous or the next speaker. Our approach was to look for frequently appearing word bigrams and trigrams including speaker names. Some of the most frequent expressions including speaker names are variants of: "I am [name]", "[name] reporting from", "[title] [name]" (Senator Dole, Prime Minister Netanyahu, correspondent Jamie Hicks), "[name] thanks", "[name] in [location]", "[name] joins us". In order to create generalized patterns, a set of 12 concept dictionaries were developed for speaker names ([name]), geographic places such as cities, countries, monuments ([location]), professions ([title]) and general communication management ([comm]) including greetings ([greet]), agreement ([agree]), acknowledgments ([thanks]), and questions [quest]). The concept dictionaries were obtained by extracting relevant items from the transcripts and then completed using additional resources such as name lists and on-line Gazetteers.

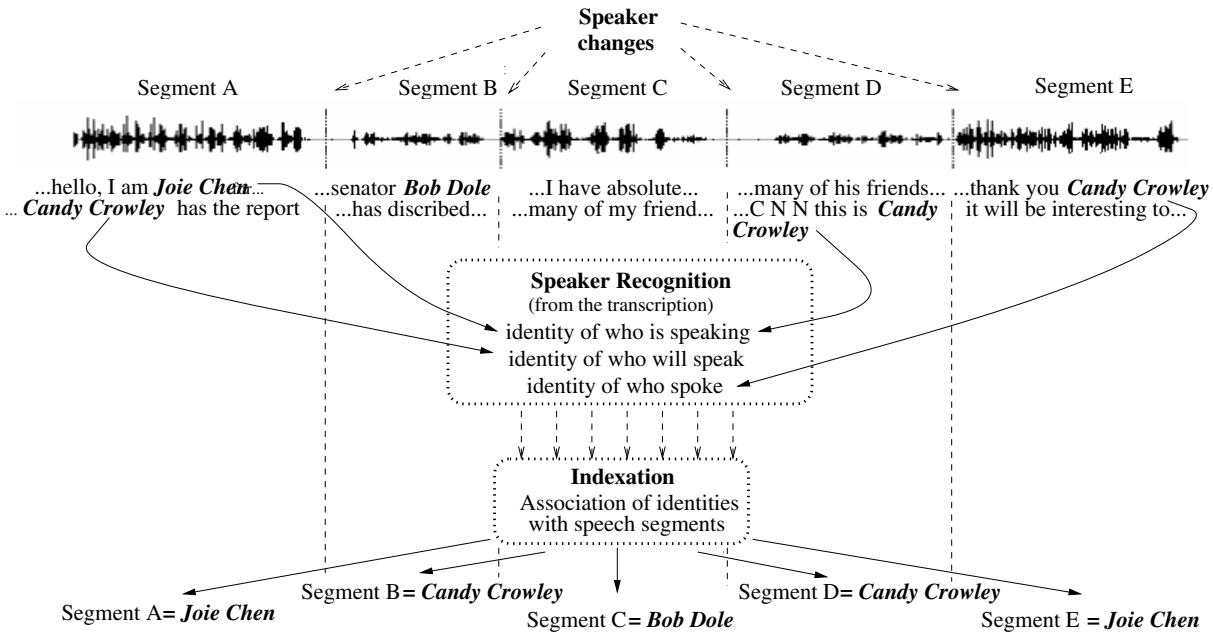

Figure 1: Example of linguistic patterns useful to identify speakers.

Figure 2: Example of parasitic information in the linguistic pattern.

Concept	*#Entries*	*#Occurences*
name	6460	22k
location	58623	17k
title	674	15k
communication	301	84k

Table 1: Concept dictionaries.

Count	*Pattern*
3162	[title] [name]
848	I_am [name]
673	[show]'s [name]
382	[agree] [name]
293	[name] [show] [location]
186	[show]'s [name] reports
176	[thanks] [name]

Table 2: Useful patterns to extract speaker identities.

Table 2 shows some of the most frequent patterns providing information about the speaker. In these examples, the speaker's name is next to the words that indicate the identity. In other cases there can be noise or parasitic information (i.e. information that does not help identify the speaker) separating the name from the trigger, as in the example in Figure 2 where information about where the speaker is located separates the "joining us" from the speaker's name. The most frequent patterns including wildcards (denoted by "$\star$") are shown in Table 3.

Count	*Pattern*
458	with [comm] us $\star$ [name]
109	joining $\star$ [name]
108	[name] $\star$ joins
45	with $\star$ [comm] me
24	[comm-agreement] $\star$ [name] reporting
24	we_are joined $\star$ by [name]

Table 3: Example patterns with wildcards ($\star$).

Upon closer examination of the most common patterns, it was observed that different patterns can be associated with the role of the speaker. For example, the show anchor will use different patterns than reporters do. The most common patterns for each of these are shown in Table 4.

3. SPEAKER DIARIZATION

The linguistic patterns have been grouped in three classes. The first group contains patterns that reveal the identity of the person who is speaking. The second group reveals who the next speaker will be and the third group indicate who just spoke.

A set of decision rules were defined to associate identities to the text segments. Before selecting which pattern applies, segments not matching any known patterns are removed. There is one rule per pattern, as well as additional rules which are used for disambiguation in certain contexts. An example requiring disambiguation is the pattern *[name] reports* $*$, where a rule

Anchor	121	[show] [name] has
	81	when we come
	51	are [comm] reading
	40	the latest on
	34	on [show] tonight
Reporter	15	is [name] reporting
	15	for [show] in
	8	update I_am +name
	8	[name] [show] the
	7	at [location] I_am

Table 4: Linguistic patterns by speaker's role.

assigns *[name]* to the next segment unless the first word of the wildcard text is *that*, in which case a filtering rule blocks the assignment.

3.1 Who is speaking: (12 patterns)

The patterns that identify who is speaking are precise and usually unambiguously reveal the identity of the speaker. The most frequently observed pattern is *I_am [name]* which occurs both at the start and the end of the program. Another common form is *This is [name]* which is often used at the end of a report.

The extracted identity is associated with the segment encompassing the pattern. For the most part the association is made directly, because the pattern is unambiguous. The most frequently matched patterns are shown in Table 5, along with the number of mistaken identities when applied to the transcriptions of the training corpus. The number of segments excluded because they are associated with unidentified speakers in the reference transcripts is given in the rightmost column. There are 2 blocking rules which filter patterns matching who is speaking. For example, a match to the pattern *[name][show]* is blocked if the left or right context matches a word in the class *[thanks]*.

Pattern	*#Matches*	*False Ident*	*Unidentified*
I am [name]	1160	1 (<0.1%)	24
[name] [show]	782	3 (0.4%)	36
this is [name]	178	5 (2.9%)	7
[name] for [show]	144	1 (0.7%)	9

Table 5: Validation of the self-speaker patterns.

3.2 Who will speak: (34 patterns)

There are a larger number of variable patterns to signal the next speaker. Some are quite clear and precise, such as *[show]'s [name] has the story* which matches *CNN'S Deborah Amos has the story* signaling the show, the speaker and the transition. This type of pattern is typical of an anchor introducing a reportage, as is *[name] [greeting]* which matches *Anne McCabe, good morning* to welcome a person calling into a talk show. A more ambiguous formulation is *[name] reports* matching *Anna Hogan reports* which can occur in different contexts, with different interpretations. This can refer to the next reporter who will speak, or can refer to a report made by some official ("the Pentagon reports", "Alan Greenspan reports"). There are 17 disambiguation rules which are applied to filter matches to the next speaker patterns.

The most frequently matching patterns are given in Table 6. If the segment is not part of an interaction, these patterns are usually unambiguous, and the subsequent segment is associated with the extracted identity. In interactions, it can be somewhat complicated to identify when the guest speaker starts speaking, since the moderator may speak for several turns after introducing the guest. At times there may be more than one guest in which case it is not always evident who is speaking first. During interactions, the linguistic pattern can be a question posed by the moderator which indicates that the next segment is a response spoken by the guest. There are a wide variety of linguistic forms characteristic of interactions that are used to maintain the communication. For interactions, a rule checks if the segment or the neighboring segment ends with a question, and if so, the identity is associated with the segment following the question. If this is not the case, a rules checks if the answer starts with an affirmation or a negation, and if so, the identity is associated with the first segment of the answer.

Pattern	*#Matches*	*False Ident*	*Unidentified*
[show] [name]	781	49 (6.8%)	65
[name] reports	431	20 (5.0%)	32
[name] has	211	32 (17.4%)	27
here's [name]	118	9 (8.1%)	7

Table 6: Validation of the next-speaker patterns.

Pattern	*#Matches*	*False Ident*	*Unidentified*
[agree][name]	244	51 (23.9%)	31
[name][thanks]	213	11 (6.1%)	32
[agree][greet][name]	128	19 (18.1%)	23
[name][agree]	40	7 (20.0%)	5

Table 7: Validation of the previous-speaker patterns.

3.3 Who just spoke: (6 patterns)

The patterns to identify the previous speaker have a precise structure. They often start with an acknowledgment, an affirmation or negation and can signal when a guest speaker has finished talking. For example, in *[gratitude] [name]* which matches *thanks Deborah Amos*, the word "thanks" serves to indicate the end of the dialog, and the name refers to the person who is not speaking.

There can be ambiguity in the linguistic event when patterns can appear both at the start or at the end of an interaction. For example, in an interview the moderator usually thanks the guest participant, but this event can happen when a guest is introduced (in which case a rule blocks the assignment of the name to the previous segment) or when the interview is finished (in this case the assignment is valid). To help disambiguate the patterns a filterning rule checks to see if the pattern matches the beginning of the final segment of an interaction. If it does, then the extracted identity is associated with the previous segment. If the pattern appears in a non-interactive context, then the extracted identity is assigned to the segment preceding the one containing the pattern. There are 5 disambiguation rules to filter matches to the previous speaker patterns. The most often matching patterns are shown in Table 7 along with the number of false identifications.

3.4 Evaluation

Table 8 summarizes the results on the complete training corpus excluding segments associated with unidentified speakers, for the three sets of patterns. The self-speaker patterns are seen to be quite reliable, and whereas higher error rates are observed when identifying the next or previous speaker. The interactive portions of the shows have more frequent exchanges between speakers, and a larger variety of lexical patterns signaling the change of speakers.

The results presented thus far were based on the training corpus from which the patterns and rules were derived. The same

Pattern	*#Matches*	*False Ident*	*Unidentified*
self-speaker	2232	28 (1.3%)	78
next-speaker	1844	210 (12.5%)	165
previous-speaker	833	181 (25%)	109
Total	4678	388 (8.9%)	335

Table 8: Speaker id error rates using linguistic patterns.

Test set	*Pattern*	*Correct*	*False Ident*	*Unidentified*
h4e97	self	40	1	0
661 segs	next	20	3	1
(294 min)	prev	7	4	0
h4e98_1	self	24	0	0
401 segs	next	14	1	0
(84 min)	prev	7	4	1
h4e98_2	self	27	0	0
428 segs	next	22	3	3
(91 min)	prev	2	3	0
h4e99_1	self	13	0	0
348 segs	next	10	7	0
(59 min)	prev	3	1	0
h4e99_2	self	16	0	0
390 segs	next	18	2	0
(90 min)	prev	4	0	1
Total		212	29	6

Table 9: The number of matching patterns and identification errors by pattern type on the evaluation data.

patterns and rules were tested on about 10 hours of unseen data from the NIST evaluation sets from 1997, 1998 and 1999. The results shown in Table 10 specify the total number of segments in each test set, the number of segments matching one of the patterns and the error rate for each pattern type. As in the validation tests, there are very few errors for the self-speaker rule. Of the 12 errors identifying the previous speaker, six arise in a potentially ambiguous context following the greeting "morning or good morning" and another 4 are due to matches of the [agree][name] pattern, where [name] refers to a third party and not the other participant in the interaction. Concerning the next speaker errors, two errors occurred on segments that preceded a portion of the audio that was not transcribed, and four errors occurred when a report began with a different speaker than was announced by the anchor (typically a report starting with a prerecorded excerpt).

Since only about 10% of the segments in the data include linguistic information which is useful for identifying the speaker, there are relatively few rule applications. However, the overall false identification rate only increases from 8.9% to 10.9%, illustrating the feasibility of the approach.

4. DISCUSSION AND CONCLUSIONS

This paper has proposed using linguistic information to associate true speaker identities with speech segments in broadcast new data. After analyzing the transcripts of the audio data, the segment types were classified as narratives or interactions, and different patterns and rules were defined for the two types of segments. In general, the linguistic patterns locate matching portions in the transcripts, but there are some problems with rules in ambiguous contexts.

A study of the speaker identification errors led to the following main causes. The linguistic patterns are located in the transcripts after normalizing names, locations, common expressions

Test	*Linguistic Patterns*			
set	*self*	*next*	*previous*	*all*
h4e97	41 (2%)	23 (13%)	11 (36%)	75 (10%)
h4e98_1	24 (0%)	15 (7%)	11 (36%)	50 (13%)
h4e98_2	27 (0%)	25 (12%)	5 (60%)	57 (11%)
h4e99_1	13 (0%)	17 (41%)	4 (25%)	34 (25%)
h4e99_2	16 (0%)	20 (10%)	4 (0%)	40 (5%)
Total	121 (0.8%)	100 (16.0%)	35 (34.2%)	256 (10.9%)

Table 10: Number of matching patterns and error rate (%) by pattern type on the evaluation data.

for communication management, via dictionaries. Some of the errors could be traced to can be proper names and places that were missing, or were present in more than one dictionary potentially causing confusions (distinguishing a person name from a city name requires knowledge of the context). In some cases the reference transcriptions are incorrect, particularly for proper names of foreign origin which may have multiple spelling variants. Occasionally there are abrupt cuts within a program, where a report is unfinished, but there is a change to another report.

The speaker identification error rate on the 150 hours of training transcripts was about 9% (1% for the current speaker, 12.5% for the next speaker, and 25% for the preceding speaker). On an independent set of 10 hours of test data, the speaker identification error rate is about 11%. While these results demonstrate the important role of linguistic information for identifying speakers, only about 10% of all segments in the corpus contain relevant patterns. Therefore this approach needs to be combined with the result of an automatic partitioning procedure which clusters segments from the same speaker, assigning within audio document identities. Then if any one of the segments has a matching linguistic pattern, the true speaker identity can be associated with all associated segments in the document. Based on the results reported here and the state-of-the-art in speaker tracking, we estimate that the correct speaker can be assigned to about 75% of the speech in broadcast news shows, without the need for training speaker-specific acoustic models. The next step will be to investigate this approach using automatic transcriptions of the audio data.

REFERENCES

[1] L. Couvreur, J.M. Boite (1999), "Speaker Tracking in Broadcast Audio Material in the framework of the THISL Project," *ESCA ETRW Accessing Information in Spoken Audio*, 84-89.

[2] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," *Eurospeech'01*, 2521-524.

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

[4] S. Johnson, "Who Spoke When? - Automatic Segmentation and Clustering for Determining Speaker Turns," *Eurospeech'99*, 2211-2214.

[5] A.E. Rosenberg, et al., "Speaker Detection in Broadcast Speech Databases," *ICSLP'98*, 202-205.

[6] D. Roy, C. Malamud, "Speaker Identification Based Text to Audio Alignment for an Audio Retrieval System," *ICASSP'97*, 1099-1102.

[7] F. Weber, L. Manganaro, B. Peskin, E. Shriberg, "Speaker Recognition and Lexical Information for Speaker Identification," *ICASSP'02*, 1099-1102.