A COMPARATIVE STUDY USING MANUAL AND AUTOMATIC TRANSCRIPTIONS FOR DIARIZATION

Leonardo Canseco, Lori Lamel, Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lcanseco,lamel,gauvain}@limsi.fr

Abstract

This paper describes recent studies on speaker diarization from automatic broadcast news transcripts. Linguistic information revealing the true names of who speaks during a broadcast (the next, the previous and the current speaker) is detected by means of linguistic patterns. In order to associate the true speaker names with the speech segments, a set of rules are defined for each pattern. Since the effectiveness of linguistic patterns for diarization depends on the quality of the transcription, the performance using automatic transcripts generated with an LVCSR system are compared with those obtained using manual transcriptions. On about 150 hours of broadcast news data (295 shows) the global ratio of false identity association is about 13% for the automatic and the manual transcripts.

1. Introduction

As the technology advances, huge amounts of information can be compacted and stored in digitized files accessible to everybody. There is a clear necessity for the users to efficiently browse, search, retrieve and access particular information from digital archives. In order to provide random access to spoken audio data, the most popular approaches have been to index and retrieve audio documents based on topics, keyword [3] or by speaker identities [5, 8]. The structuring of an audio document into acoustically homogeneous segments according to the speaker identity and the background and channel conditions is known as acoustic diarization [1, 7].

This study aims to explore the significance of linguistic information in the diarization process. 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 main types of information allows a broadcast news audio recording to be structured into individual news stories for further diarization.

In order to identify weakness and strengths of a linguistically-based diarization 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. The LIMSI Large Vocabulary Continuous Speech Recognition (LVCSR) system has been used to produce the automatic transcriptions [4]. The recovered set of experiences and observations emphasizes the important role of the linguistic information in a diarization process. A previous study reported in [2] contains details about the proposed approach along with results using manually produced transcripts. The general process for linguistic-based diarization is summarized in the next section.

2. Linguistic-based Diarization

The typical broadcast news show has an anchor who leads the program, usually introducing the reporters, the show's guests and the upcoming commercial breaks. The reporters take part in the broadcast when a report starts, which can be done on-site or could have been previously recorded. The sequence of events reveals the structure of the show, as depicted in Figure 1. From this example, it is clear that speaker introductions occur frequently in the broadcast, appearing in the linguistic patterns and let listeners know who will be the next person to speak.

Our approach is based on observing the most frequent appearing word bigrams and trigrams including speaker names. Recurrent words surrounding a speaker identity correspond to locations, professions, shows names and general communication management including greetings, agreements, acknowledgments, questions and responses. In order to generate generalized patterns, 12 dictionaries were created from the recurrent words. These were constructed by extracting relevant items from the transcripts and complemented with additional resources such as name lists and on-line Gazetteers. In the first two columns of Table 1, the most relevant dictionaries are shown along with their corresponding number of entries.

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, who will speak or who just spoke. For each one of these situations rules are defined to associate an identity with a speech segment corresponding to a *speaker turn*. For the patterns that reveal who is speaking the extracted identity is associated with the speech segment encompassing the pattern, this is defined as a "self-speaker" rule. For those patterns which announce who will speak, the recovered identity is associated with next speech segment, defining a "next-speaker" rule. For patterns matching who just spoke, the speaker's name identified is related with previous speech segment, defining a "previous-speaker" rule. Some additional disambiguation and blocking rules are required to limit the identity association rules, details of the complete process can be found in [2].

0-7803-9479-8/05/$20.00 © 2005 IEEE

ASRU 2005

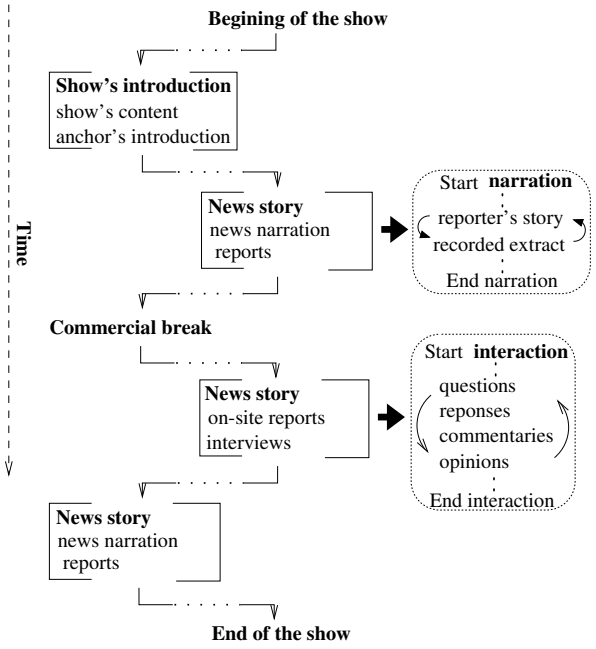

Figure 1: Example of a broadcast news program sequence with story segments from announcer, interviews and reports

		#Matches	
Concept	*#Entries*	*Manual*	*Auto*
name	6460	26469	26678
location	58623	20618	22415
title	674	15787	16642
communication	301	92279	70089
show name	14	4715	6794

Table 1: Concept dictionary coverage on the *Hub4-E* manual transcriptions (Manual) and on the automatic transcriptions (Auto).

3. Transcriptions

In this study the English Broadcast News Speech *Hub4-E* corpora distributed by LDC were used for development purposes. In total there are about 150 hours of broadcast news data from 295 news shows broadcast from 1993 to 1998. The data 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). This collection of programs contains a large amount of speech manually segmented into *speaker turns* with close to perfect orthographic transcriptions. These data were used to identify linguistic patterns for diarization and to validate the rules. The 1997, 1998 and 1999 DARPA/NIST evaluation test sets are used to assess the approach on about 10 hours of unseen data. More detail about the manual and automatically generated transcriptions used for evaluation purposes are as follows:

Manual transcriptions contain the speech segmented into accurate *speaker turns*, each *turn* containing time markers of beginning and end of utterance, as well as the true speaker name of who spoke (when it is possible); for the speakers who are not known distinct identifiers are used (i.e., spkr1, janedoe1). These transcriptions represent an ideal case for developing a linguistic approach due to accurate transcriptions and a correct speech into *speaker turns*.

Automatic transcriptions of the news shows have been generated with a version of the LIMSI LVCSR system used in DARPA evaluations for broadcast news (BN) speech transcription [4]. The system uses continuous density hidden Markov models with Gaussian mixture for acoustic modeling, and *n-gram* statistics estimated on large text corpora for language modeling. Since the standard Hub4-E acoustic training data is being used in this work as development data for the diarization procedure, new acoustic and language models were trained for this task without using any of the manually transcribed BN data. The acoustic models were trained on about 140 hours of data from the *TDT2* corpus using a lightly supervised approach [6], and the language model was estimated on about 1 billion of words of texts. The word error rate on this data is around 20%. The system also performs a segmentation and classification of the signal according to different acoustic conditions (bandwidth, background noise). This audio partitioning process segments the speech into a series of estimated *speaker turns*, avoiding problems caused by linguistic discontinuity at speaker changes [4].

4. Diarization results

As a first validation step, the coverage of each dictionary was quantified on the manual and automatic transcriptions. Table 1 compares the matched items in both transcriptions, showing that the number of matched items in the automatic transcripts is slightly higher than in manual transcripts for all concept dictionaries with the exception of the communication management dictionary. The differences can be attributed to automatic transcription errors and to ambiguities between some speaker names and location names. Since the training data are not manually tagged, we can only indirectly assess which items have been correctly transcribed and tagged by looking at the effectiveness of the matched patterns.

The reliability of the self, next and previous-speaker rules was then quantified by aligning the hypothesized *speaker turns* with those in the reference transcripts, and comparing the associated identities with the reference ones. As can be supposed, the effectiveness of the patterns is correlated with the transcription accuracy. Recognition errors cause mismatches in the patterns, for example some names are replaced by another name or are only partially transcribed.

Inexactly estimated *speaker turns* generated by the automatic partitioning process can also mislead the identity association rules. We observed that some of the estimated *turns* contain speech from more than one speaker. This poses problems in evaluating the validity of the assigned identity to an impure *speaker turn*; how much of this association is valid? or how much of the assigned identity is right or wrong? In terms of speech alignment, this can be quantified by measurements of cluster purity and cluster coverage [4]. In terms of pattern accuracy these are not strictly speaking errors since the pattern correctly matches one of the referenced identities associated with the speech segment. We therefore defined the following identity association cases: ***C1:*** The extracted identity is associated with a pure *speaker turn* and it matches completely the reference identity (first and last name). ***C2:*** the extracted identity is associated with an impure *speaker turn* and it matches completely one of the reference identities in this segment. ***C3:*** the extracted identity is associated with a pure *speaker turn* and

Evaluation	*Manual Transcription*			*Automatic Transcription*		
Cases	*self-spkr*	*next-spkr*	*prev-spkr*	*self-spkr*	*next-spkr*	*prev-spkr*
#C1	2137 (98.6%)	1186 (73.6%)	135 (18.4%)	1239 (75.3%)	756 (64.2%)	85 (20.2%)
#C2	-	-	-	75 (4.5%)	73 (6.2%)	6 (1.4%)
#C3	28 (1.3%)	209 (13.0%)	390 (53.2%)	231 (14.1%)	150 (12.7%)	154 (36.7%)
#C4	-	-	-	18 (1.0%)	10 (0.9%)	10 (2.3%)
#False id	4 (0.1%)	217 (13.4%)	208 (28.4%)	84 (5.1%)	188 (16.0%)	165 (39.3%)
#undef.	81	146	119	73	111	74
Total Matches	2250	1758	852	1720	1288	494

Table 2: Diarization rates using linguistic patterns on manual and automatic *Hub4-E* transcriptions.

Evaluation	*Manual Transcriptions*			*Automatic Transcription*		
Cases	*self-spkr*	*next-spkr*	*prev-spkr*	*self-spkr*	*next-spkr*	*prev-spkr*
#C1	115 (95.0%)	50 (55.0%)	7 (16.0%)	94 (84.0%)	38 (60.3%)	8 (21.0%)
#C2	-	-	-	2 (1.7%)	3 (4.8%)	-
#C3	7 (5.0%)	22 (24.8%)	18 (40.9%)	7 (6.2%)	10 (15.9%)	11 (29.0%)
#C4	-	-	-	-	-	-
#False id	-	16 (20.2%)	19 (43.1%)	9 (8.0%)	12 (19.0%)	19 (50.0%)
#undef.	-	3	1	-	2	1
Total Matches	122	91	45	112	65	39

Table 3: Diarization rates using linguistic patterns on manual and automatic transcripts (*97-98-99 Hub4-E* evaluation data).

it partially matches the reference identity. ***C4:*** the extracted identity is associated with an impure *speaker turn* and it partially matches one of the reference identities. ***Undef:*** the pattern matches an undefined speaker in the reference (these cases are identified and excluded in the evaluation results). ***False Id:*** None of the above conditions apply, so this is an erroneous identity association with a *speaker turn.*

Table 2 summarizes the performance of the self-speaker, next-speaker and previous-speaker rules when these are applied to manual and automatic transcriptions of the *Hub4-E* corpus. The same trends can be seen for both transcripts. 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 found in the automatic transcription is about 18% below the number obtained using the manual transcriptions. The total number of false identity associations (for all three rules), represents of about 9% of the total number of identity associations for both the manual and automatic transcriptions.

The results in Table 3 specify the total number of identity associations for same linguistic patterns and rules when applied to 10 hours of unseen data from the NIST evaluation sets from 1997, 1998 and 1999. As for the development data (*Hub4-E*) there are few errors for the self-speaker rule and the other rules are have the same tendencies. The total number of identity associations decreases by about 20% for the automatic transcription compared to the manual one. The total percentage of false identity associations of the three rules is about 13% and 18% for the manual and automatic transcriptions respectively. Therefore, the percentage of false associations is similar for the two sets of transcriptions.

5. Story Dynamics

With the aim of associating identities to portions that are not covered by the three speaker rule types, an approach based on the role of the speakers and on the types of news program has been developed. The example in Figure 1 illustrates how the dynamics of a news program can be classified.

The dynamics of the news stories allow a portion of a show to be structured as a function of the role of those taking part in the segment. The process consists of identifying the type of the news story (e.g. interactive or narrative), and then as a second step associating a speaker role to the speech segments (when possible). A final step relates the roles with the true speaker names.This process is described in the following:

5.1. Speaker role recognition

Since information about the role of the speakers is not included in the reference transcripts, it was defined empirically. Four roles were identified for broadcast news programs: anchor, reporter, guest and announcer. Table 4 classifies the most frequent patterns which indicate the role of the speakers. These patterns can include wildcards (denoted by "*"). The labels in the table represent speaker names by "[name]", news sources by "[show]", and acknowledgments by "[thanks]". Geographic places are denoted by the "[loca]" label, the upcoming commercial breaks are represented by the "[break]" label, and the demonstrative pronouns by the "[dem]" label (see [2] for more details).

Within the anchor class, there are self-introduction patterns spoken exclusively by the anchor as well as anchor introductions expressed by the announcer. For example, the most frequent pattern is "[dem][show] * I am [name]" which matches "This is ABC Night Line I am Joie Chen". There are a total of 11 patterns related to the anchor role. From these patterns, the name of the anchor is identified.

The reporter class groups together patterns with a precise structure used by the anchor to introduce reporters. The most frequent pattern corresponds to "[show][name]" which matches "CNN's John Zarella". There are 20 patterns of this type from which it is possible to determine the names of the reporters.

The guest class groups together patterns which may be spoken by the anchor or by a reporter. These correspond to speaker introductions and speaker acknowledgments. The most frequent guest introduction has the form "joining * is [name]" which

Role	*Pattern*	*#Matches*
Anchor	[dem][show] * I am [name]	253
	[greet] * Iam [name]	169
	[break] * I am [name]	102
	[dem][show] * with [name]	57
Reporter	[show] * [name]	781
	[name] reports	431
	[name] has	211
	here's [name]	118
Guest	joining * is [name]	1247
	[thanks][name]	238
	[name] * joins	112
	joined * [name]	58
Announcer	[dem][show] * with [name]	57
	from [show][dem][show]	29
	[dem][show]*[loca][name]	23
	from [loca][demo][show]	22

Table 4: Patterns for the speaker role identification.

matches "joining us this evening is Michael". And the most frequent guest acknowledgment corresponds to "[thanks][name]". There are 8 patterns in this class which allow the identity of guests to be determined.

The announcer is identified by very precise patterns, which are expressed exclusively by the program's announcer. The linguistic message in these patterns is very limited. The most frequent patterns contain the program name followed by an anchor introduction: "[dem][show] * with [name]" which match "this is world news tonight with Peter Jennings". There are 9 patterns which characterize an announcer, which also allow the name of the anchor to be extracted.

5.2. News story classification

The linguistic styles found in broadcast news shows can be broadly classified in two major categories: portions that are narratives and portions that are interactives. These are illustrated by the narration and interactive boxes in Figure 1.

Interactive portions are primarily interviews, where there is an exchange of ideas and often with explicit questions and answers. The beginnings and ends of interactions are automatically detected when the same speaker name appears in a speaker introduction pattern ("Senator Bob Dole is joining us") and in a speaker acknowledgment pattern ("thank you senator Dole"). Interactions are classified by the number of guests (extracted from the guest introductions) and by the number of show hosts (extracted at the beginning of the show). Interview are modeled as follows: *speaker turns* containing explicit questions are related to the moderator (the host) and the speaker turns starting with affirmations or reflections are attributed to the guests. However, reflections are attributed to the guests only if the previous *speaker turn* contains an explicit question. Table 5 shows the most frequent self-expressions (e.g. affirmations and reflections) and explicit questions.

Narratives are characterized by anchors or reporters who typically present the news in third person singular or plural. Speech extracts from public figures which are played as part of the report are often expressed in first person singular or plural form. The beginning and end of a narrative are automatically detected when the same speaker name appears in a reporter introduction

#Matches	*Self-expr.*	*#Matches*	*Questions*
2664	I think	396	do you
1693	You know	204	who was
613	I don't	167	how much
439	Of course	143	what was

Table 5: The most common linguistic self-expressions and questions in the corpus.

pattern ("N.P.R's Melissa Block reports") and in a sign-off report pattern ("Melissa Block N.P.R news, New York"). This style of communication allows associations for narrative portions to be hypothesized as follows: *Speaker turns* containing self-expressions or patterns of first person (i.e. "I think") are related to speech extracts. *Speaker turns* in the third person ("Senator Dole said"), excluding self-expressions, are associated to reporters and anchors.

When unambiguous, the anchor's name is associated to the moderator *turns*, and the guest's name to the guest *turns*.

6. Story Dynamics Results

Tables 6 and 7 report the diarization performance as a function of the story dynamics on the development and evaluation data respectively. For interviews, the moderators are denoted as "*m*" and the guests as "*g*"; and narratives are labeled as such. In the development data 5 program types are observed, with the vast majority being narratives which account for about 80% of stories. The main problem for the interactions is to assign the correct moderator name to the *speaker turns*when there are 2 moderators. For interviews with 2 guests and 2 anchors, the false identity association is very high (60% for 6 shows with manual transcripts). For interviews with one guest in shows with one anchor lower error rates of around 20% are obtained. For narratives, lower false association rates (of around 5%) are reported for both manual and automatic transcriptions, on the evaluation data. The frequent occurrences of narratives in broadcast news data explains the high number of identities associated to the these portions.

7. Conclusions

This study has highlighted the importance that linguistic information can have for diarization of broadcast news data. True identities can be determined from transcripts of the broadcasts and associated with *speaker turns*. The linguistic approach to speaker diarization makes use of frequent word sequences which include speaker names, revealing the identity of the current, the next or the previous speakers. Linguistic-based diarization was evaluated on the ideal case using a manual transcription of what was said along with manual speaker turn segmentations, as well as on automatic transcriptions with estimated *speaker turns*.

Similar tendencies for false identities associated to the *turns* were observed for both conditions. The self-speaker rules are quite reliable, with the lowest error rates of 0.1% for the manual transcripts and 5% for the automatic ones. This rule is also the most frequent in the data, as reporters usually explicitly sign on and off. The highest false associations are for the previous-speaker rules, which about 38% error on the manual transcripts, and 40% on the automatic ones. As described in [2] there is more ambiguity in the patterns and rules to detect the previous-

	Dynamics	*#C1 (%)*	*#C2 (%)*	*#C3 (%)*	*#C4 (%)*	*#False Id*	*#Undef.*
	1m-1g	189 (82.1%)	-	-	-	41 (17.9%)	57
	1m-2g	142 (78.4%)	-	-	-	39 (21.6%)	37
Manual	*2m-1g*	138 (55.0%)	-	4 (1.5%)	-	109 (43.5%)	13
	2m-2g	6 (37.5%)	-	-	-	10 (62.5%)	-
	narrative	1753 (98.2%)	-	-	-	33 (1.8%)	144
	1m-1g	64 (55.7%)	11 (9.5%)	1 (0.8%)	-	39 (34.0%)	5
	1m-2g	26 (50.0%)	1 (2.0%)	16 (30.7%)	-	9 (17.3%)	10
Automatic	*2m-1g*	66 (32.9%)	10 (5.0%)	14 (7.0%)	3 (1.4%)	108 (53.7%)	14
	2m-2g	2 (14.2%)	3 (21.4%)	2 (14.3%)	1 (7.1%)	6 (42.9%)	5
	narrative	882 (85.9%)	52 (5.0%)	27 (2.6%)	2 (0.1%)	64 (6.3%)	88

Table 6: Diarization using story evaluated on the manual (top) and on the automatic (bottom) transcriptions for the development data (*Hub4-E*). *m* represents a moderator and *g* a guest.

Dynamics	*#C1 (%)*	*#C3 (%)*	*#False Id*	*#Undef.*
1m-2g	10 (45.5%)	10 (45.5%)	2 (9.0%)	-
narrative	61 (74.4%)	17 (20.7%)	4 (4.9%)	2
1m-2g	6 (35.3%)	8 (47%)	3 (8.2%)	-
2m-2g	4 (100.0%)	-	-	1
narrative	49 (84.5%)	6 (10.3%)	3 (5.2%)	1

Table 7: Diarization using story dynamics on manual (top) and automatic (bottom) transcripts (*eval97-98-99* data. *m* represents a moderator and *g* a guest.

speaker than for the other two cases. The total number of false identity associations for the sets of rules is about about 9% of the total number of associations for the 150 hours of development data (*Hub4-E*) using manual transcripts and 12% with automtic ones. On the evaluation test sets these are about 13% and 18% for the manual and automatic transcription respectively.

Interactive portions of the broadcast news programs were observed to have higher false association rates. Therefore an analysis of the story dynamics in the data was carried out with the aim of classifying the data into single-speaker reports (narratives) and multi-person interactions. Narratives are able to be detected quite reliably, and are more frequent than interactions, covering a much larger proportion of the speech data. The false identity association is about 5% for narratives. Initial results for interviews involving multiple persons are mixed (false ids ranging from 20-50%), in part due to the limited number of test samples for the various conditions.

It has been shown that linguistic information can be used to enrich the information in automatic diarization broadcast news data. This information can be extracted from automatic transcriptions via the use of linguistic patterns developed using manual transcripts. Investigation of the combined use of linguistic and acoustic information for speaker diarization is currently underway. The linguistic information can potentially help resolve situations where data from a given speaker is split into multiple clusters (usually representing different acoustic environments).

8. References

[1] Barras, C., Zhu, X., Meignier, S. and Gauvain., J.L., "Improving Speaker Diarization." In Proc. DARPA RT04, Palisades NY, November 2004.

[2] Canseco L., Lamel L., and Gauvain, J.L., "Speaker Diarization From Speech Transcripts", Proc. International Conference on Spoken Language Processing, pp. 2004.

[3] Doddington, G., "Speaker recognition based on idiolectal differences between speakers," *Eurospeech'01*, pp. 2521-524, 2001.

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

[5] Jin, H., Kubala, F., and Schwartz, R., "Automatic Speaker Clustering", Proceedings of the Speech Recognition Workshop, pp. 108-111, 1997.

[6] Lamel, L., Gauvain, J.-L., and Adda, G., "Lightly Supervised and Unsupervised Acoustic Model Training", in Computer Speech and Language IDEAL, Vol. 16, pp. 115-129, 2002.

[7] Reynolds, D., and Torres-Carrasquillo, P., "Approaches and Applications of Audio Diarization." In Proc. IEEE ICASSP, Philadelphia, March 2005.

[8] Roy, D., and Malamud, C., "Speaker Identification Based Text to Audio Alignment for an Audio Retrieval System.", In Proc. ICASSP, pp. 1099-1102, 1997.