LEVERAGING SIDE INFORMATION FOR SPEAKER IDENTIFICATION WITH THE ENRON CONVERSATIONAL TELEPHONE SPEECH COLLECTION

Ning Gao,[1,2] *Gregory Sell,*[2] *Douglas W. Oard,*[1,2] *Mark Dredze*[2]

[1]College of Information Studies and UMIACS, University of Maryland, College Park
[2]Human Language Technology Center of Excellence, The Johns Hopkins University

ABSTRACT

Speaker identification experiments typically focus on acoustic signals, but conversational speech often occurs in settings where additional useful side information may be available. This paper introduces a new distributable speaker identification test collection based on recorded telephone calls of Enron energy traders. Experiments with these recordings demonstrate that social network features and recording channel metadata can be used to reduce error rates in speaker identification below that achieved using acoustic evidence alone. Social network features from the parallel Enron email collection (37 of the 41 speakers in the telephone recordings sent or received emails in the collection) improve speaker identification, as do social network features computed using lightly supervised techniques to estimate a social network from more than one thousand unlabeled recordings.

Index Terms— Speaker Identification

1. INTRODUCTION

Understanding conversational speech is a challenging task with many potential applications; examples include providing access to recorded meetings, making sense of the panoply of records that can be generated in lifelogging, and—our focus in this paper—analysis of telephone conversations that were recorded for regulatory compliance purposes. While some practical applications can present additional challenges (e.g., acoustic conditions such as additive noise or room reverberation), specific applications can also provide new opportunities (e.g., the availability of side information that characterizes the broader context in which the conversations occurred). In this paper we focus on some of those opportunities.

Collections that are representative of the situated use of conversational speech in specific conditions have been available for some time. Notably, the AMI and AMIDA projects [1] created a corpus of meeting recordings consisting of two types of meetings: a design scenario, and naturally occurring meetings in a range of domains. Side information in those cases includes email messages between the meeting participants. A new collection of Mission Control Center conversations from NASA's Apollo Program is expected to be released soon, and in that case the side information consists of metadata indicating the roles and expected participants on specific intercom circuits plus thousands of written documents (e.g., technical reports) [2]. Lifelogging is also of interest to speech researchers [3, 4, 5], although the first public lifelogging test collection (from NTCIR) is focused on image rather than spoken content [**?**, 6]. Despite the potential for collections like these to be used to explore contextual features, the research community as a whole has initially pushed hardest on fully exploiting the acoustic information common to all of these applications.

In this paper, we describe experiments with a conversational telephone speech collection for which five types of side information are available. The collection is built from recorded phone conversations made by or to Enron energy traders, and the task that we study is closed set identification of known speakers. Our principal focus is on coupling acoustic evidence with specific types of side information to improve performance on this speaker identification task.

Our basic approach is as follows. For each phone recording, we first use a conventional speaker identification system trained only on acoustic evidence to rank each of the candidate speakers according to the probability of their being one of the speakers in a specific call. Then we use one of five types of side information (two types of social network features, two types of channel features, and detected mentions of known name variants) to re-rank the speaker candidate list for each call. Our social network features are based on the contact frequency between candidate speakers, which can be observed directly in the email collection, or which must be estimated for the automatic (and thus potentially erroneous) predictions for telephone calls that are initially based on acoustic evidence. These features are then used to re-rank the candidate speakers in a manner that puts speakers who are more likely to speak with highly ranked speakers closer to the top of the list. The channel features are learned from a list (that had been manually prepared independently for a court case) of the channels on which specific speakers might be expected or are estimated using acoustic evidence for a set of calls with unknown participants. We can then re-rank the candidate speaker list for any recording in a way that promotes people who would be expected to appear on the known channel of each call. Finally, as with any collection we can also use the content of the call to re-rank candidate speakers in ways that promote people who might be referred to by some name mentioned during the call. Our experiments with name mentions are more preliminary, focusing on whether further improvements can be obtained using name variants (e.g., nicknames) that are learned from the associated email collection.

In the remainder of this paper, we first introduce the Enron Conversational Telephone Speech (CTS) collection and the associated side information in Section 2. Section 3 then introduces our evaluation measures. Baseline results using acoustic evidence alone are presented in Section 4, followed by results using five types of contextual features in Section 5. Our discussion of those results in Section 6 is then followed by brief mention of future work in Section 7.

2. TEST COLLECTION

In this section, we describe the full Enron CTS and Email corpora in Section 2.1 and 2.2, followed by details of the test collection that we have built from those corpora in Section 2.3.

2.1. Enron CTS Collection

The Enron CTS collection consists of 1,731 recordings that together total 47.8 hours. Each recording includes at least one, and sometimes more than one, telephone conversation made from or to an Enron energy trader. These phone recordings were made for regulatory purposes, and were posted to the Internet by the Snohomish County Public Utility District in Oregon pursuant to their use in a lawsuit.[1] For those recordings that include more than one call, the calls are typically separated by some combination of dial tone, Dual-Tone Multi-Frequency (DTMF) dialing codes, and ring tone. Transcripts were manually prepared for 57 of these recordings for use in court. Those transcripts are available as scanned page images, for which Optical Character Recognition (OCR) yields a low character error rate. We therefore use uncorrected OCR when using these transcripts for content representation (for some of our experiments.) The transcripts include the channel (which we manually correlated to the channel metadata for those 57 recordings) start time, and duration. Speaker turns in the transcripts are labeled with the name of the speaker when that speaker could be reliably identified by the transcriber; on average there are 1.4 identified speakers per manually transcribed recording (which on average includes 1.5 calls). The document that contains the transcripts also contains a table showing which speakers were either frequently or sometimes observed on each channel. That table had been manually prepared for use in the court case; we transcribed it manually for use in our experiments. For each recording file, its recorded channel can be extracted from the file name (e.g., audio file SNO-163-1-20000803-9102233-9163302.wav is recorded from channel 163).

2.2. The Enron Email Collection

The Enron email collection distributed by Carnegie Mellon University (CMU) [7] has been widely studied by researchers with interests in social network analysis or automated processing of informal text, but it has been little used in speech processing research. The collection contains 517,431 unique messages saved by the owners of 152 email accounts. The collection contains about 133,000 unique email addresses, which after clustering (by associated full names) yielded entity models for 124,475 unique person entities [8]. Elsayed processed this collection in two ways that are useful to us: performing deduplication (resulting in 248,573 unique messages) and constructing a collection-specific knowledge base of person entities [8]. Each entity model contains all known email addresses and all known name variants for that person, where the name variants (e.g., first name or nickname) were learned from the salutations (e.g., "Hi Dave!") or signatures (e.g., "Thanks – Jim") of message body text in the collection, or inferred from automatic tokenization of full names or email addresses found in the headers of those messages. Each entity model also includes counts for messages sent to or receive from each other entity. Because the collection was built from messages stored by or on behalf of individuals, some of whom retained messages longer than others, these contact statistics are a convenience sample of the actual contact statistics for Enron as a whole. Notably, the 152 email accounts used to build the collection are naturally substantially over-represented in comparison with other Enron accounts.

[1] https://web.archive.org/web/20050206035158/http://www.enrontapes.com/files.html

2.3. Building a Speaker Identification Test Collection

We partitioned the 57 recordings for which manual transcripts are available into a training set containing 28 recordings and a test set containing 29 recordings. Across the 57 recordings there are a total of 41 different speakers whose names are available from the transcripts, and we were able to manually associate 37 of these names with the full names of people represented in Elasyed's collection-specific knowledge base. We selected 28 of the 57 recordings as a training set in a manner that ensured that the full set of 41 known speakers would each be represented in at least one training recording. For each of these 28 recordings in the training set we then manually diarized segments for each known speaker for use in training the speaker models. We use the remaining 29 transcribed recordings as the basis for our test set. We manually segmented this test set into individual telephone calls; this results in a total of 45 test calls (some of which are very short). The known speakers for each call were then manually determined from the scanned transcripts, which had been manually prepared.

3. EVALUATION MEASURES

Our core task is to identify which speakers from the training set are present in a call from the test set. There is always at least one known speaker, and often there are two. There are calls with three or more speakers, but none include more than two known speakers. In the experiments that follow, we consider this problem from two different paradigms: detection and identification. Detection is the task of finding the recordings that include a particular speaker, while identification is determining which of the known speakers is in the particular recording-of-interest. In both cases, we consider the problem as closed-set, meaning that we are concerned with identifying only the known speakers, and at least one known speaker is present in every call in the 45-call test set.

First, to evaluate detection, we measure performance using the detection cost function (DCF) [9], a common metric in speaker recognition.

$$DCF = C_M P_M P_T + C_{FA} P_{FA} (1 - P_T) \quad (1)$$

The cost of misses and false alarms, C_M and C_{FA} respectively, are both set to 1, and the prior probability P_T is set to 0.03, matching the evaluation prior to maximize smoothness. This metric involves pooling all scores for every speaker-call pairing and then setting a universal detection threshold. We report scores at the oracle threshold, yielding the minimum DCF for the system at that prior.

Second, to evaluate identification, we compute classification error, which simply determines how often the correct speaker is given the highest score for a particular file. In the cases where there are two known speakers in the file, the lower-scored known speaker is considered correct if his/her score is higher than all incorrect speakers. Given that there are 41 possible speakers, classification error for chance decisions would be roughly 0.976 (in actuality, it is slight lower, due to the instances with two correct speakers).

Third, for additional insight into identification performance, we supplement with an evaluation measure based on mean reciprocal rank (MRR), common in information retrieval. The fundamental statistic that we seek to estimate is the rank of each known speaker in the list of scores for a particular recording. When there are multiple known speakers, we replicate the system's ranked list with one of the two known speakers removed in each. We then take the harmonic

mean of the rank of the known speaker in each such list:

$$R = \frac{n}{\sum_{i=1}^{n} \frac{1}{r_i}} - 1, \quad (2)$$

where r_i is the rank of the known speaker in list i and n is the number of lists (i.e., the number of speaker-call pairs). R can thus be interpreted as the number of rank positions below what would otherwise be a perfect ranking at which the system places the correct speaker. In our experiments R is always between zero and one, but in principle R is unbounded and R for a random ranking would be about 20.

For each of our measures, lower values are preferred, with zero being the lowest possible value. Moreover, because zero indicates perfect performance (corresponding to consistently putting every known speaker at the earliest possible rank), our measures are all ratio measures (i.e., measures in which, for example, a value twice as large is twice as bad); this makes small differences meaningful.

4. ACOUSTIC SPEAKER IDENTIFICATION

Our i-vector [10] baseline system for speaker identification uses acoustic evidence alone to rank all known speakers. The system used the Fisher English corpus to train the universal background model (UBM) and total variability (T) matrix, and NIST SRE '04, '05, '06, and '08 data to train the Probabilistic Linear Discriminant Analysis (PLDA). The resulting system performs competitively with other state-of-the-art acoustic-only i-vector systems on standard evaluations. Each trial resulted in a log likelihood ratio that was calibrated via unsupervised methods [11] using the 1,674 unlabeled recordings (1,731 minus the 57 labeled training or test recordings).

In this case, the recordings typically include multiple speakers, so diarization was required in front of the i-vector system. For each file, a diarization system based on i-vector segments [12] was used to estimate the number of speakers and the bounding marks for each. These marks were scored against the model in question, and then fed into the recognition system, resulting in an i-vector for each estimated speaker. The maximum score among those was retained as the overall score for the model in that conversation.

Since we are considering the tasks here as closed-set, we also applied Bayes' rule to the scores, normalizing each model's score by the sum of all model scores for that test file. Note that this is an improper operation when more than one true speaker is present in the file, and modifying the process to account for this could be valuable future work.

The baseline system achieves a DCF of 0.67, classification error of 0.56, and a (harmonic) expected rank R of 0.73 (Table 1).

5. RE-RANKING TECHNIQUES

In this section, we discuss our approaches for improving over the acoustic baseline with side information. We re-rank speaker scores using a social network (Section 5.1), channel information (Section 5.2), or name variant detection (Section 5.3).

5.1. Social Network Re-Ranking

Our first experiment uses a social network to re-rank speakers; we begin with the email social network. We have 41 known speakers, 37 of whom sent or received email in the CMU Enron email collection. For each of these 37 speakers, we know from Elsayed's knowledge base how often they communicated with each of the other 36 known speakers in the email collection. Conceptually, then, we could reasonably expect a conversation to more often involve frequent communicants than rare ones.

We formalize this as follows. If two known speakers were present in the same email header (i.e., if one sent and the other received an email message, or if both received the same message) we build an edge between them in the social network, and we set the weight of that edge to be the frequency with which they communicate. Let e_l denote the sum of the edge weights that are connected with one of the speakers (which we refer to as the left speaker), e_r to denote the sum of the edge weights that are connected with the other ("right") speaker, e_{lr} to denote the (undirected) edge weight between the left speaker and the right speaker, and $\sum e$ to denote the sum of all the edge weights in the social network. The score of a pair is then calculated as:

$$s_p = \frac{1}{2}\left(\left(1 + \frac{e_l}{\sum e}\right)s_l + \left(1 + \frac{e_r}{\sum e}\right)s_r\right)\left(1 + \frac{e_{lr}}{\sum e}\right). \quad (3)$$

The equation shows the five factors that influence our estimate of whether the left and right speakers are true speakers in the conversation: the acoustic score s_l of the left speaker, boosted by the degree to which the left speaker is a frequent communicant ($\frac{e_l}{\sum e}$); the acoustic score s_r of the right speaker, boosted by the degree to which the right speaker is a frequent communicant ($\frac{e_r}{\sum e}$); and a boosting factor applied to both that reflects the degree to which these two speakers communicate with each other ($\frac{e_{lr}}{\sum e}$). The use of two individual boosting factors is a precision-oriented design reflecting that only frequent communicants with high acoustic ranks have the power to "pull" up other speakers. We then re-rank the speakers according to their highest associated s_p (or their original score in the case of speakers with no observed pairs). Table 2 illustrates the ranking by acoustic score, the pair ranking, and the final re-ranked list using an actual example from the collection (with names anonymized). The first pair places *Speaker01* and *Speaker03* on the re-ranked list, in that order; the second pair then results in addition of *Speaker04*, and the final insertion of speakers missing from any pair adds *Speaker02*.

If we knew which speakers had actually participated in some large number of phone calls, we could apply a similar process to leverage the telephone social network, but true labels are only known for a small number of phone calls. Instead, we use our acoustic baseline system described in Section 4 to predict which speakers participated in each of the 1,703 non-training recordings (1,731 minus the 28 labeled training recordings). By counting these predicted telephone interactions, we can generate a similar network to that drawn from the emails, thus producing an alternative re-ranking that we can evaluate to determine whether the larger size and more accurate observability in the email social network yields better results than the smaller and less accurately estimated, but perhaps more highly comparable, telephone social network.

Our measures provide different insights in this case, shown in Table 1. DCF shows a dip in performance for both networks, with the email network hurting the score slightly less (from a baseline of 0.67 to 0.72) than the phone network (0.74). Classification error, by contrast, improves in both cases, again with the email network (0.49) yielding slightly better results (from a baseline of 0.56 to 0.49) than the telephone network (0.51). In terms of R, the telephone social network turns out to be the clear winner, however, improving by 11% relative to the baseline (from 0.73 to 0.65) compared to 0.70 for the email network. It seems clear from these results that both networks help for identification, but not necessarily for detection, and the measures disagree about which network is more useful.

Single Source	DCF@0.03	Classification Error	R
Baseline	0.67	0.56	0.73
Email Social Network	0.72	0.49	0.70
Phone Social Network	0.74	0.51	0.65
Manual Channel	0.69	0.43	0.53
Estimated Channel	0.76	0.46	0.57
Name Variants	0.43	0.21	0.17
Multiple Sources	**DCF@0.03**	**Classification Error**	**R**
Email Social Network & Estimated Channel	0.74	0.48	0.55
Phone Social Network & Estimated Channel	0.69	0.51	0.64
Email Social Network & Manual Channel	0.66	0.43	0.50
Phone Social Network & Manual Channel	0.61	0.46	0.56

Table 1. Evaluation of the Re-ranking results.

Acoustic Rank	Ranking of Speaker Pairs	Final Re-ranked List
Speaker01	Speaker01 & Speaker03	Speaker01
Speaker02	Speaker04 & Speaker01	Speaker03
Speaker03		Speaker04
Speaker04		Speaker02

Table 2. A re-ranking example. The true speakers are *Speaker01* and *Speaker03*.

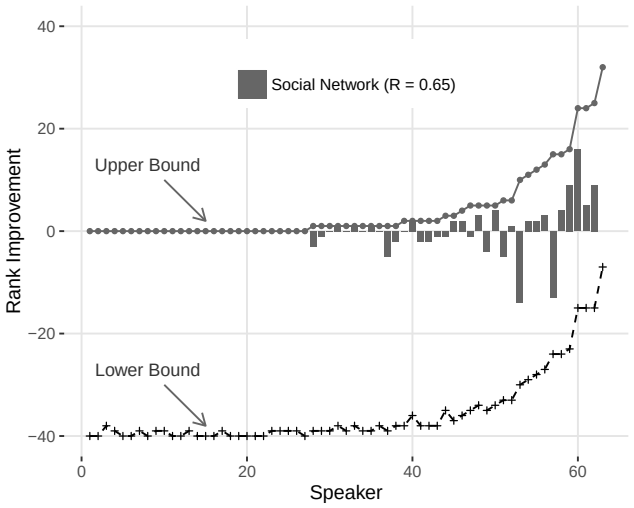

Fig. 1. Rank improvement of true speaker after re-ranking.

Aggregate results can mask important insights, so Figure 1 provides a compact visualization of where this approach works, and where it fails, for the self-trained telephone social network. In this plot, the Y axis shows the change in rank of the true speakers as a result of the side information for each test trial, which is itemized on the x-axis and sorted by the initial R ranking. The upper and lower bounds of possible rank changes are also shown for context.

As can be seen, no speaker that the acoustic evidence had initially correctly placed at the best possible rank (i.e, no speaker for which the upper bound on the possible improvement was zero) was adversely affected by re-ranking. Notably, four speakers (each of which started out near the top of the list) achieved the maximum possible improvement. Re-ranking resulted in more changes—both positive and negative—for speakers lower in the list, moving the rank up in 17 cases and down in only 12.

Speaker	Main Channel(s)	Other Channel(s)
Speaker05	1, 13	2, 3, 14, 15
Speaker06	26	16, 25, 51

Table 3. Example channel information for speakers.

5.2. Channel Re-Ranking

The Enron CTS collection also includes metadata indicating on which channel each call was recorded, as well as a list (prepared professionally for use in a lawsuit) that indicates which people were typically recorded on which channels. Table 3 shows an anonymized excerpt from this list. The "Main Channel(s)" are those on which the compiler of the list expected to see the speaker most often, whereas "Other Channels(s)" are those on which they chose to note that the speaker was also sometimes present. Some channels repeat as main channels for different speakers, suggesting that there was some sharing of phones (e.g., during different work shifts), meaning that this channel information is not sufficient on its own for predicting the true speaker. It is easy to see how we might use this information to re-rank the speakers, since if we know that *Speaker05*'s main channel is channel 1 and that channel 1 is not *Speaker06*'s main channel, then *Speaker05* may be a better speaker candidate than *Speaker06* when the call is recorded on channel 1.

For comparison with this manually compiled channel information, we also tried a process similar to that used to build our telephone social network to estimate channel probabilities for each speaker on the 1,703 non-training recordings. To do this, we used the observed channel mappings learned from acoustic evidence and the training recordings to estimate how often each speaker was likely to be recorded on each channel.

We then formalize the re-ranking process as follows. Let $T = (t_1, \ldots, t_m)$ be the m unique channels on which recordings in the collection have been recorded, and $F_c = (f_1, \ldots, f_m)$ be the number of calls in which candidate speaker c was detected using each channel based on acoustic evidence. We then calculate a new score $s_c^{'}$ for each candidate c based on the acoustic prediction score s_c and an estimate of the probability that speaker c is observed on channel i:

$$s_c^{'} = \left(1 + \frac{\lambda f_i}{\sum_{q=1}^{m} f_q}\right) s_c, \tag{4}$$

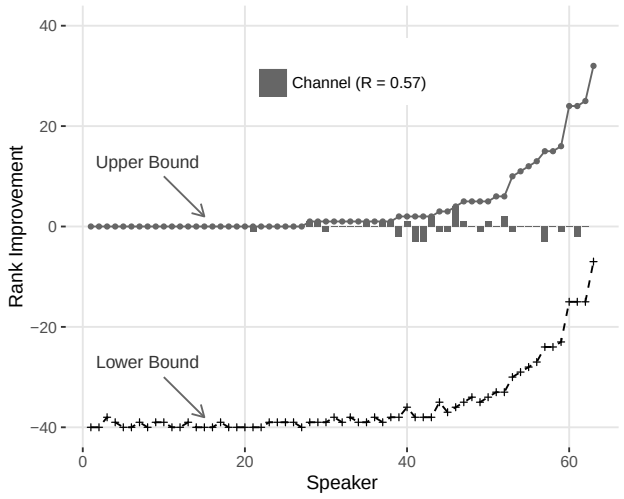

Fig. 2. Rank improvement using the estimated speaker-channel information.

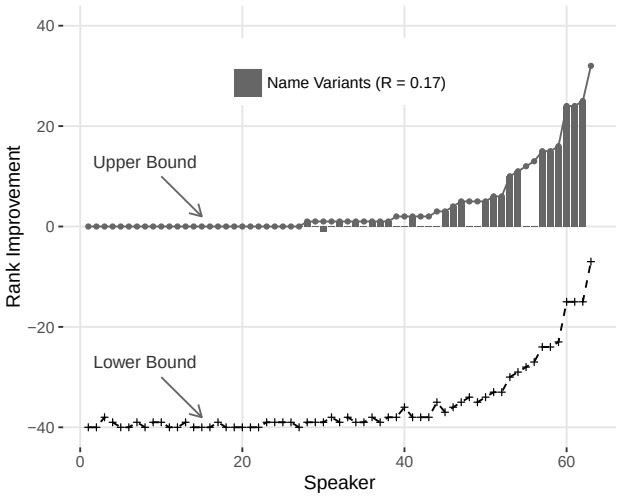

Fig. 3. Rank improvements by using the named variants automatically detected in manual transcripts.

where λ is a parameter to adjust for the relative weight of the channel information. In our experiments, we arbitrarily set $\lambda = 1$. This is a simple approach to explore the contributions of this type of data, but a more principled estimation and incorporation, such as in [13], would be a valuable extension in the future.

To use the same re-ranking process with the manually prepared list, we arbitrarily set the number of calls to 2 for main channels, to 1 for other channels, and and 0 for channels that are not listed. Although this process is not optimized, it serves as a useful reference to compare against the results of our automated estimates that are estimated from a larger, but noisier, set of examples from what amounts to semi-supervised training.

When using the manually prepared speaker-channel table, detection performance as measured by DCF degrades slightly (from 0.67 to 0.69), while the estimated channel information degrades DCF more substantially (from 0.67 to 0.76). However, both classification error and R show improvements in identification performance with either source of channel information. Using the manually prepared table improves classification error (from 0.56 to 0.43) and R (from 0.73 to 0.53), while using the automatic channel estimates improves classification error somewhat less (to 0.46) and R correspondingly less (to 0.57). This improvement in R from the fully automated technique is a 22% relative improvement that is significant under a two-tailed paired t-test (at p $<$0.05).

Figure 2 shows a compact analysis of the case-by-case results for the provided speaker-channel table that is structured identically to that in Figure 1. In this case there are 11 improvements and 12 reductions in rank, but many of the improvements are near the top of the ranked list and at or near the upper bound, whereas the reductions in rank occur only for correct candidates that were already at or below rank 3, and they come nowhere near the lower bound. Classification error and R both reward these improvements more than they penalize reductions in rank with those characteristics.

5.3. Name Mention Re-Ranking

Frequently speakers will identify themselves at the beginning of a conversation (e.g., "Snohomish, Jay." "Hey Jay, Holly."). Since we know (from the knowledge base constructed from the email collection) how 37 of our 41 speakers might be referenced, we can easily make use of evidence from named mentions. We formalize this as follows. For each speaker candidate, we first automatically match it to at most one person entity in the knowledge base introduced in Section 2.2, and we then make a list of all name variants associated with that entity. The knowledge base includes information about the frequency with which each variant was observed, so we record that in our list as well. For example, an individual could be mentioned as *John* in the Enron email collection five times, and mentioned one time as *Johnny* (nickname). Therefore, we estimate the probability of that person entity being mentioned as *John* as $p = 5/6 = 0.83$. We automatically scan the first two speaker turns in the manual transcript of each call in the test set for all name variants in our list, and then we rescore each candidate as:

$$s_c^{'} = s_c * (1 + \beta p), \tag{5}$$

where $s_c^{'}$ and s_c are defined above, β is a parameter that could be tuned to adjust the weight of name variant evidence (set to 1 in our experiments), and p is the estimated probability from the knowledge base that the candidate is mentioned by that name (0.83 in the example).

This approach is something of a special case of phonotactic speaker recognition, which recognizes speakers by the content of their speech. However, the key difference here is the data itself. In collecting data for other sets, such as NIST SRE, self-identification is typically removed from the audio. This data, on the other hand, more closely reflects real-life interactions. The approach proves to be effective, substantially improving DCF from 0.67 to 0.43, classification error from 0.56 to 0.21, and R from 0.73 to 0.17. Figure 3 shows that nearly every candidate whose name was detected in the first turn of each speaker turned out to be the true speaker. However, the use of nicknames learned from email body salutations and signatures (e.g., Johnny as a nickname for John) has only a small effect; when we remove those nicknames from the knowledge base and use only the provided first and last names, performance degrades only very slightly (DCF from 0.43 to 0.44, classification error from 0.21 to 0.23, and R from 0.17 to 0.18).

We must, however, offer two caveats regarding this experiment. First, we utilized manual transcripts for these experiments, and the

degree to which this result can be replicated using speech recognition will depend on the ability of that speech recognition system to detect the name mentions. Task-specific tuning of the language model might help with that, since the list of name variants is available in advance. Second, there is little ambiguity in the name variants among our set of 41 speakers (only 37 of which have associated knowledge base entries). With far larger speaker sets, effective techniques for disambiguation would become important. Results from entity linking in email indicate that this is an entirely tractable problem (when social network evidence and evidence from content are used together) [14, 15], but of course both the social network and the content evidence are generally less accurately observable in speech than in email text.

5.4. Combination of Multiple Sources

Table 1 also shows the effect of fusing the re-rankings with simple score summation. For these experiments, we only explored the combinations of one type of social network with one type of channel information, yielding four fusion pairs. The DCF detection measure again tells a different story to the two identification measures. When combined with manual channel information, some improvement over the acoustic baseline is observed with the telephone social network (from 0.67 to 0.61), whereas only a slight difference in DCF is observed when manual channel information is combined with evidence from the social network (0.66). Combinations with estimated channels yield no improvements (and indeed slight degradation) in DCF. The identification measure R improves when the email social network is used together with channel information, when compared to the already-good results for channel information (both for manual channel information from 0.53 to 0.50) and for estimated channel information (from 0.57 to 0.55). No similar improvement is seen from using the telephone social network. The classification error measure shows no improvement over either type of channel information for either type of social network.

6. DISCUSSION

In considering the collective results of all the above experiments, there are a few overall impressions. First, by either of the identification measures (classification error or R) it is clear that the incorporation of social network evidence helps a little and that channel information helps somewhat more. The use of nicknames learned from the email collection doesn't help much at all since in our test collection there is only one person "Stanley" referred to by a nickname "Stan". However, we might find greater usefulness of nickname matching when there are more people involved in a larger data collection. Among the two social networks we tried, using the evidence from communication patterns in the email network results in consistent improvements, both with and without the complementary evidence from channel information. The telephone social network is even more helpful than the email network when used alone, but when used in combination with the channel information from either source the telephone social network yields no further improvement over using channel information alone. One plausible explanation for this is that all channel information, manual or automatic, ultimately relies on acoustic evidence, and acoustic evidence also informs our estimate of the telephone social network. When combining evidence, the email social network is thus a better choice as a complementary source of evidence.

However, the results also show that the same side information is much less helpful for speaker detection, at least as characterized by our DCF measure, and, in fact, it is often detrimental. This is likely due to one of two causes: (1) the side information is affecting test files differently, which causes the scores to pool poorly; or (2) the side information is causing more false alarms, which are weighted more heavily in DCF at the selected prior. The latter option appears to be more likely based on examination for Figures 1 and 2, where we see that the true speaker is moved down the list almost as often as it is moved up. DCF measures this balance differently than does classification error or R, both of which are more strongly influenced by what happens at (or in the case of R, near) the top rank. This explanation is also supported by the fact that the name variant side almost never hurts the rank of the true speaker in Figure 3, and this was the only side information to improve DCF on its own.

The fact that channel information consistently outperformed social network information as a side information feature both by classification error and by R is intriguing, but we must note that the structure of this test collection (with telephone lines used by specific people being recorded) is particularly well suited to the use of channel features. In other applications (e.g., cases in which trunk lines are recorded) it may be social network features that are of greater use. Our results do show, however that we are able to estimate channel assignments from acoustic evidence sufficiently reliably to be useful, and to achieve results close to what manual annotation was able to perform.

One final observation is that both of our automatically-derived sources of information (the telephone social network and the estimated channel information) offer the promise for a double benefit from future improvements to acoustic speaker recognition techniques, since both automatically derived sources leverage acoustic speaker recognition. So not only will the acoustic baseline improve, but better estimates will be made on the unlabeled data, possibly resulting in better side information as well.

7. CONCLUSION

We have introduced a new speaker identification test collection and explored five approaches for incorporating side information to improve performance on a speaker identification task. We have illustrated how the Enron conversational telephone speech collection can be used for such experiments, and we have used that data to demonstrate that automatic predictions can be used as a basis for social network and channel analysis to improve speaker identification. Our experiments with name mention features using manual transcripts yielded improvements that were unsurprising, but they allowed us to study the effect of adding nicknames to the set of known name variants.

In Section 5.4, we explore the combination of different types of side information by simply summing the scores. One future direction would be applying a machine learning framework to learn the appropriate combination model for the collection. Another direction that we would like to explore would be to experiment with the use of spoken term detection for person names, thus automating a process that our present experiments with manual transcripts have shown to have substantial potential for yielding improvements.

8. ACKNOWLEDGEMENTS

This work has been supported in part by NSF Grants 1065250 and 1618695 and by a Mellon Foundation Coherence at Scale Doctoral Fellowship. Opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the views of NSF or the Mellon Foundation.

9. REFERENCES

[1] Steve Renals, Thomas Hain, and Hervé Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2007, pp. 238–247.

[2] Douglas Oard, Abhijeet Sangwan, and John HL Hansen, "Reconstruction of apollo mission control center activity," in *the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH*, 2013.

[3] Alexander G Hauptmann, Jiang Gao, Rong Yan, Yanjun Qi, Jie Yang, and Howard D Wactlar, "Automated analysis of nursing home observations," *IEEE Pervasive Computing*, vol. 3, no. 2, pp. 15–21, 2004.

[4] Ali Ziaei, Abhijeet Sangwan, and John HL Hansen, "Prof-life-log: Personal interaction analysis for naturalistic audio streams," in *IEEE Workshop on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7770–7774.

[5] Abhijeet Sangwan, Ali Ziaei, and John HL Hansen, "ProfLifeLog: Environmental analysis and keyword recognition for naturalistic daily audio streams," in *IEEE Workshop on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4941–4944.

[6] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal, "Overview of NTCIR-12 lifelog task," in *Proceedings of NTCIR-12 Conference*, 2016, pp. 354–360.

[7] Bryan Klimt and Yiming Yang, "The enron corpus: A new dataset for email classification research," 2004, pp. 217–226, Springer.

[8] Tamer Elsayed and Douglas W Oard, "Modeling identity in archival collections of email: A preliminary study.," in *CEAS*, 2006, pp. 95–103.

[9] "The NIST Year 2010 Speaker Recognition Evaluation Plan," (Available at `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`), 2010.

[10] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–98, 2011.

[11] Niko Brümmer and Daniel Garcia-Romero, "Generative modelling for unsupervised score calibration," in *IEEE Workshop on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1680–1684.

[12] Gregory Sell and Daniel Garcia-Romero, "Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014, pp. 413–417.

[13] Alan McCree, "Estimating and Exploiting Language Distributions of Unlabeled Data," in *Proceedings of Interspeech*, 2010, pp. 209–214.

[14] Ning Gao, Mark Dredze, and Douglas W. Oard, "Person entity linking in email with nil detection," *Journal of the Association for Information Science and Technology*, vol. 68, no. 10, pp. 2412–2424, 2017.

[15] Ning Gao, Douglas Oard, and Mark Dredze, "Support for interactive identification of mentioned entities in conversational speech," in *International Conference on Research and Development in Information Retrieval (SIGIR)*, 2017, pp. 953–956.