FINDING SPEAKER IDENTITIES WITH A CONDITIONAL MAXIMUM ENTROPY MODEL

Chengyuan Ma[1]*, Patrick Nguyen*[2]*, and Milind Mahajan*[2]

[1] School of Elec. Engineering
Georgia Institute of Technology
cyma@ece.gatech.edu

[2] Speech Research Group
Microsoft Research
{panguyen,milindm}@microsoft.com

ABSTRACT

In this paper, we address the task of identifying the speakers by name in audio content. Identification of speakers by name helps to improve the readability of the transcript and also provides additional meta-data which can help in finding the audio content of interest. We present a conditional maximum entropy (*maxent*) framework for this problem which yields superior performance and lends itself well to incorporating different types of information. We take advantage of this property of maxent to explore new features for this task. We show that supplementing standard lexical triggers with information such as speaker gender and position of speaker name mentions afford us large gains in performance. At 95% precision, we increase the recall to 67% from the trigger baseline of 38%.

Index Terms— Speaker recognition, Maximum entropy methods

1. INTRODUCTION

The amount of audio/video content available online has increased tremendously over the recent years. This has generated considerable interest in providing effective tools to the users for finding the content of interest and then effectively navigating through it. Automatic speech recognition (ASR) and audio indexing play an important role in making the content searchable. However, additional meta-data for the content such as speaker turn segmentation is very useful in improving the readability and understanding of the ASR output. The task of segmenting the audio content by speaker turns is known as speaker diarization. Significant recent work on automatic speaker diarization has taken place recently [1]. Speaker diarization identifies all the audio segments belonging to a particular speaker however it does not identify the speaker by name. Ability to identify the speaker would be further beneficial to the user since the real name of the speaker allows the user to quickly bring to bear all the real world knowledge about the person into understanding the content. The user may be better able to understand the context and assign varying importance to different speakers.

Speaker identification by name provides useful meta-data for audio search as well since it allows the user to search by the speaker name in addition to the text of what was said. The task of identifying the speaker names has been aptly called "who really spoke when?" in [2] and this is the task that we will address in this paper.

In general, there are several useful pieces of information which could help for this task. In a formal news, an informational show or a presentation, the speakers may introduce either themselves or other speakers by name; or they may simply refer to each other by name while talking. If we have heard some speakers previously and know them by name, we could use the acoustic similarity to identify these same speakers in the content of interest. There may be some meta-data associated with the content such as show description which names the speakers. Finally, visual features such as visual similarity or captions in video could be used if the video is available as well.

Previously, [2, 3, 4] have developed systems which used only the linguistic patterns such as word n-grams in the audio transcripts for identifying the speakers. [5] has used acoustic similarity based on acoustic models for known speakers for whom training data exists to identify their re-occurrence in the test data.

Our approach builds upon this previous work and uses a combination of linguistic and acoustic information rather than relying solely on either. We propose the use of conditional maximum entropy framework (*maxent*) for this task. Conditional maximum entropy framework provides a sound theoretical mechanism for integrating different types of features. We show how it can be used to combine linguistic and acoustic features which have been used previously. We also investigate new features for this task in this framework and show its benefit empirically. In this paper, we have limited ourselves to acoustic and linguistic information sources and have not explored the use of meta-data or video information.

This paper is organized as follows. In Section 2, we frame the speaker identification problem using a simple probabilistic model. In Section 3, we provide an overview of the current state-of-the-art system. In Section 4, we describe our conditional maximum entropy model framework. We then proceed describe the novel features we investigated in Section 5. In Section 6, we present the experimental results on the Hub4 Broadcast News corpus and conclude in Section 7.

2. PROBLEM FORMULATION

The input for our problem consists of a unit of acoustic data (such as a single show) with associated text transcript and segmentation by speaker turns (diarization). Input speaker turns are grouped by the same speaker, such that all turns of a single speaker, in that unit of acoustic data, form a speaker cluster c. The process of obtaining the text transcript and diarization is out of the scope of the problem we address. Our task consists of assigning a speaker name s to each speaker cluster c. We use a conditional probability model $P(s|c)$ and conditional maximum probability as the decision rule. In some usage scenarios, the cost of assigning a wrong speaker name could be higher than the cost of not assigning a speaker name at all. To deal with this, we add a threshold ϑ on the conditional probability to the decision rule. So, we assign a speaker name $\hat{s} = \arg\max_s P(s|c)$ if $P(\hat{s}|c) \geq \vartheta$. Otherwise, we are not confident enough to make a prediction and leave the speaker name as unknown. We use preci-

All work done while Chengyuan Ma was an intern in Microsoft Research.

sion and recall as the evaluation metrics and changing ϑ allows us to control the trade-off between the two. The particular operating point for ϑ is usually driven by the application. We will describe further details and refinements of this general model in the following sections.

3. OVERVIEW OF STATE-OF-THE-ART

At the time of writing, the state of the art is a probabilistic trigger feature system described by Tranter [2], which we henceforth refer to as the Tranter system. In this section, we provide a high-level overview of the Tranter system. Please refer to [2] for details of the system and a detailed discussion of the evaluation framework.

3.1. Lexical rule learning

The idea is to latch on linguistic patterns which are characteristic of a speaker denomination.

In Broadcast News, speakers may be announced, introduce themselves, or have their name reminded by the next speaker. Therefore, a speaker name for an audio segment may be found in the previous, current, or next segment. For instance, after encountering a training utterance "This is John Simpson speaking from ...", a 5-gram rule: `Current Speaker: This is [name] speaking from` is formed. A probability is associated with this rule which is a relative frequency over the training data of how often the `[name]` is the correct speaker of the current segment in this context. Similarly, rules are formed for predicting if the `[name]` refers to the speaker of the previous or the next segment. All possible 2-5 gram rules are formed using 2 tokens to the left and right of the `[name]` token. The Tranter system maps certain groups of words to category tokens to increase robustness. For example, both "BBC News" and "ABC News" are replaced by category token `[SHOW]`.

Since the number of rules thus collected may be prohibitively large and the relative frequencies may not be good estimates for the sparsely occurring rules, the Tranter system only uses rules which have 5 or more instances in the training data. The Tranter system also uses a threshold on the probability associated with each rule to control the number of rules which are retained. Varying this threshold can also affect the trade-off between precision and recall.

3.2. Lexical rule application

At test time, to determine the probability of assigning a speaker name s to a cluster c, first the set of all applicable rules $\mathcal{R}(s)$ which support such assignment is determined. In determining the set of rules $\mathcal{R}(s)$, a back-off system is used to retain only those rules which are not completely contained by a larger n-gram rule. For example, the rule `Current Speaker: is [name] speaking` will cause the smaller rule `Current Speaker: is [name]` to be dropped from $\mathcal{R}(s)$. After determining $\mathcal{R}(s)$, Tranter system uses a heuristic combination of associated rule probabilities p_r as follows:

$$P(s|c) \propto 1 - \prod_{r \in \mathcal{R}(s)} (1 - p_r) \tag{1}$$

Only those speakers for which $\mathcal{R}(s)$ is non-empty are considered as the candidate speaker names for the cluster c. Thus, candidate speaker set for a given cluster c are all the speaker names which occur at least once either in the acoustic segments belonging to the cluster or in the acoustic segments which are adjacent to the acoustic segments of the cluster (due to the presence of Next Speaker and Previous Speaker rules). We use the notation $\mathcal{S}(c)$ to refer to the set of candidate speakers for cluster c.

3.3. Lexical features for maxent

We have reproduced the Tranter system as our baseline. We used only ϑ to control the trade-off between precision and recall. We have referred to our reproduction of the Tranter system as the N-gram system in the remainder. We have tested the N-gram system on some of the experimental configurations reported in [2] and found the results to be very similar.

We have also used exactly the same information used by the Tranter system in the maxent framework. In maxent, the model depends on the data only through the feature functions. We define a feature function corresponding to each rule in the N-gram system. For instance, the feature function corresponding to the rule: `Current Speaker: is [name] speaking` will be 1 whenever the speaker name which is preceded by *is* and followed by *speaking* is the hypothesized speaker for the speaker cluster under consideration. This system with just these feature functions constitutes the baseline for exploration of new features in the maxent framework and is referred to simply as *maxent*. We note that unlike the N-gram system, in maxent framework the parameters are trained jointly and the parameter combination is a normal part of the training process.

4. MAXIMUM ENTROPY MODEL

The first contribution of this work is to replace the probabilistic trigger model with a regularized log-linear model [6]. Specifically, the probability of a hypothesized speaker is given by:

$$P(s|c) = \frac{1}{Z_\lambda(c)} \exp\left(\sum_{k=1}^{F} \lambda_k f_k(s,c)\right). \tag{2}$$

Here, f_k are feature functions which together form the feature vector of size F. The λ_k are the parameters of the maxent model and $Z_\lambda(c)$ is a normalization constant which ensures that probabilities sum to one. Specifically, $Z_\lambda(c) = \sum_{s' \in \mathcal{S}(c)} \exp(\sum_{k=1}^{F} \lambda_k f_k(s', c))$. The model parameters λ_k are trained using Generalized Iterative Scaling to optimize the posterior probability of the parameters λ given the training data with a Gaussian prior [7].

We use a single pooled variance σ^2 for all features, set heuristically to 1.5. In practice, system performance seemed relatively insensitive over a large range of σ. Since the normalization constant tends to contrast the ground truth label against any other label, this model is best suited for classification tasks such as this one. Also, it is ideally suited for combining features of a different type. Multiple features may fire simultaneously without a special combination rule.

Continuous features are quantized and split into several features, thereby providing the ability to learn arbitrarily shaped histograms.

Finally, we remark that the Tranter combination rule follows a log-linear functional form for $1 - P(s|c)$ with binary features for each rule r, and $\lambda_r = \log(1 - p_r)$.

5. NEW FEATURES

The second contribution of this work lies in the investigation of new features for this task. Tranter lexical features are always present in the system and constitute the baseline for work on new features. We describe below additional features which we investigated.

5.1. Position information

Intuitively, the name of the speaker is more likely to be mentioned around the first or last appearance of the speaker in a show. We pursue this intuition by introducing features related to the relative position of the segment containing the claimed speaker name with respect to the first and the last segment of the speaker cluster. The first segment of the cluster is represented as S_f and the last segment as S_l.

For each cluster, we defined seven relative position categories, depending on whether the segment in which claimed speaker name occurs is *1)* just before S_f, *2)* in S_f, *3)* just after S_f, *4)* just before S_l, *5)* in S_l, *6)* just after S_l, *7)* anywhere else. We associate a binary feature with each position category which is 1 if the claimed speaker name occurs within that category and is 0 otherwise.

Based on the analysis of the training set, we found that the position feature is informative and decided to use it. Figure 1 shows the discrimination power of the position feature on the test set.

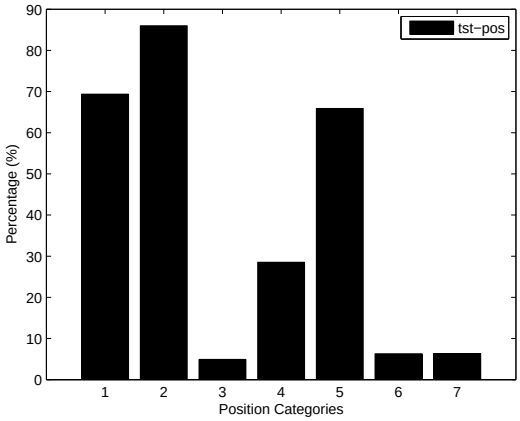

Fig. 1. Discrimination potential of position category: Bar for each position represents the percentage of claimed speaker names in the position which are true speaker names. Wide variation in the bar heights is indicative of the useful discrimination power.

In contrast, we performed similar analysis over the training set for the position of the claimed speaker name within its own segment and found that it did not seem to possess any discriminative power and therefore did not use it.

5.2. Gender information

Another effective feature pertains to gender. The intuition here is that the gender as indicated by the name of the claimed speaker should match the gender which is inferred from the acoustic data for the speaker cluster of interest. The first name is used to infer the gender of the claimed speaker name. The gender is also extracted from the audio of the speaker cluster using a gender-dependent GMM with 192 Gaussians each trained on the training set. The binary feature matches the two independently acquired gender identities.

5.3. Acoustic speaker ID

We use the information contained in the normalized cross log-likelihood value which is widely used in speaker identification literature [8]. This is useful only for speakers which are common between the train and test sets. We train speaker adapted 192 Gaussian GMM acoustic models for all speakers in the training database. Let k be a training speaker with associated audio O_k, and maximum *a posteriori* adapted model M_k. The presented test speaker j is defined similarly on test data. The universal gender dependent background model is M. We compute the normalized cross log-likelihood (CLR) as:

$$\text{CLR}(k,j) = \frac{1}{N_k}\log\frac{p(O_k|M_j)}{p(O_k|M)} + \frac{1}{N_j}\log\frac{p(O_j|M_k)}{p(O_j|M)}. \qquad (3)$$

If greater than zero, it is indicative that k and j might be different instances of the same speaker. The histogram of normalized CLR values on the test data in Figure 2 supports this hypothesis. To convert the CLR value into features we quantized the CLR values into 4 buckets and associated a binary feature with each bucket. The values were selected by looking at the histogram on the training set.

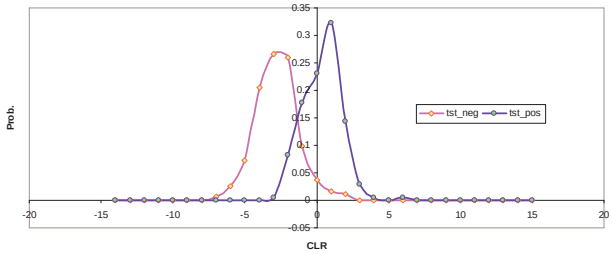

Fig. 2. CLR Histogram on test (tst). Pos line represents CLR histogram for same speaker, and neg line represents CLR histogram for different speakers. Significant non-overlapping areas of the histograms indicate usefulness.

6. EXPERIMENTS

Experiments were carried out on two subsets of the Hub4 Broadcast News training database (train96: LDC97S44, LDC97T22; train97: LDC98S71, LDC98T28). To increase the statistical significance of results, we decided to use 85h of test.

6.1. Experimental framework

The Hub4-1997 was used as the training data while the Hub4-1996 is the test set. Both sets have been labelled manually with reference transcriptions and speaker names. In this paper, we have used reference (manual) text transcript and reference diarization for the test set. The approach is extensible to the use of ASR generated text transcript and the use of automatic diarization system.

The training set is about 83 hours and the test set is about 85 hours. The number of speaker clusters for which true speaker identity is available is 1267 for the the training and 1296 for the test set. Similar to the Tranter system, our approach can recover the speaker name only if the speaker name occurs in the transcript of the speaker cluster or neighboring segments. This imposes an upper bound on maximum achievable recall on the test set of 83%.

Similar to the Tranter system, only the multiple word speaker names are treated as true speaker identities and all the others are ignored. We are aware of some errors in the speaker name labeling from [2] as well as our own observations, however, we have not corrected these corpus errors and we treat the misspelled names as if they were different speakers.

6.2. Results and discussion

Figure 3 shows the results of all the system configurations on the test set. With the same lexical trigger features, the maxent system outperforms the N-gram system. We attribute the gain to the discriminative nature of the conditional training of parameters in the maxent framework. In contrast, N-gram system uses maximum likelihood estimates of parameters. The maxent system jointly trains the feature combination parameters in contrast to the N-gram system where the rule combination is heuristic.

With the position feature and gender information incorporated into the maxent system, system performance is much improved. The results also show that the acoustic feature (CLR) bears little incidence on overall performance. As mentioned before, CLR can only help for the common speakers in the train and test sets. In our experimental setup, there are about 150 such common speakers. These common speakers account for 10% of the speakers in the test set and 30% of the test set in time-weighted proportion. An insufficient number of bins, or the relatively low discriminative power of the CLR might be additional reasons for the lack of performance improvement.

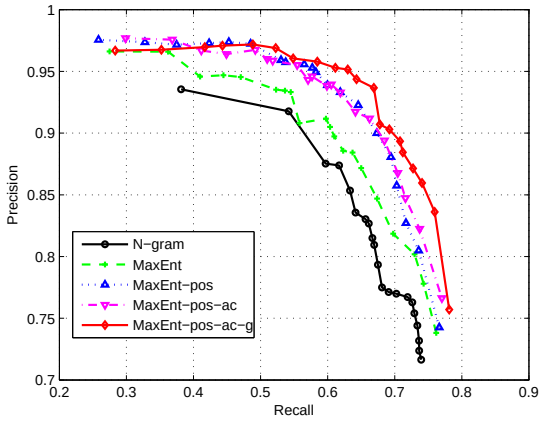

Fig. 3. Precision-Recall trade-off curve: N-gram and maxent use identical lexical N-gram features, maxent also shown with the addition of position (pos), acoustic (ac), and gender (g) features. Shift in the curve to the upper-right corner indicated improved overall performance.

Since many applications would want to keep the false alarm rate low, looking at the recall values at high precision levels (say 95%) is another alternative way of analyzing system performance. We use this method to compare all systems in Table 1. Note that since the N-gram system does not reach 95% in our experiments, we use the recall corresponding to the maximum achievable precision level of 93.5% for that system. This likely over-estimates the reported performance of the N-gram system by a small amount.

7. CONCLUSIONS AND FUTURE WORK

We have presented a conditional maximum entropy framework for identifying speaker names in audio content. This framework has the advantage of providing a flexible way to combine different types of features. We have successfully taken advantage of this flexibility. To our knowledge, this is the first attempt to combine lexical and acoustic information for this task. We also introduced new features such as position and gender features and empirically verified their usefulness for this task. We showed the benefit of the sound method for joint training of the parameters in the maximum entropy framework by demonstrating better performance than the state-of-the-art N-gram system while using the same set of lexical features. Our experimental results show that at a fixed precision of 95%, our best maxent system increases the recall from 38% to 67%.

Table 1. System recall scores at 95% precision.

System	Recall
N-gram	38%
maxent	47%
maxent + pos	58%
maxent + pos + ac	58%
maxent + pos + ac + gender	67%

In this paper, we have used the reference transcripts and diarization for the test set. In the future, we will study how replacing these two components by their automatic counterparts will affect the performance. Other factors which can affect the performance are the use of an automatic named entity recognizer and a test domain which is different from the broadcast news domain. We are also planning further study on syntactical, acoustic and prosodic features.

8. REFERENCES

[1] S. Tranter and D. A. Reynolds, "An overview of automatic speaker diarisation," *IEEE Trans. on Speech and Audio Processing*, To appear.

[2] S. E. Tranter, "Who really spoke when? Finding speaker turns and identities in broadcast news audio," in *Proc. ICASSP*, 2006, pp. 1013–1016.

[3] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcriptions," in *Proc. ICSLP*, 2004, pp. 1272–1275.

[4] L. Canseco-Rodriguez, J.-L. Gauvain, and L. Lamel, "Towards using STT for Broadcast News Speaker Diarization," in *DARPA RT04 workshop, Palisades, NY*, November 2004.

[5] D. Moraru, L. Besacier, and E. Castelli, "Using *a priori* information for speaker diarization," in *A Speaker Odyssey: The Speaker Recognition Workshop*, 2004.

[6] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[7] S. Chen and R. Rosenfeld, "A survey of smoothing techniques for ME models," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 37–50, January 2000.

[8] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communications*, vol. 17, pp. 91–108, August 1995.

[9] W. Antoni, C. Fredouille, and J.-F. Bonastre, "On the use of linguistic information for broadcast news speaker tracking," in *Proc. ICASSP*, 2006, pp. 1021–1024.