Speaker Recognition in a Multi-Speaker Environment

Alvin F Martin, Mark A. Przybocki

National Institute of Standards and Technology
Gaithersburg, MD 20899 USA
alvin.martin@nist.gov, mark.przybocki@nist.gov

Abstract

We discuss the multi-speaker tasks of detection, tracking, and segmentation of speakers as included in recent NIST Speaker Recognition Evaluations. We consider how performance for the two-speaker detection task is related to that for the corresponding one-speaker task. We examine the effects of target speaker speech duration and the gender mix within test segments on results for these tasks. We also relate performance results for the tracking and segmentation tasks, and look at factors affecting segmentation performance.

1 Introduction

Speaker recognition has generally been viewed as a problem of verifying or recognizing a particular speaker in a segment of speech spoken by a single speaker. But for some applications of interest the problem is to verify or recognize particular speakers in a segment of speech in which multiple speakers are present. Automatic systems need to be able to segment the speech among the speakers present and/or to determine whether speech by a particular speaker is present and where in the segment this speech occurs.

During the past three years the annual NIST Speaker Recognition Evaluations (see [1] and [2]) have included tasks that address these issues in addition to the conventional task of speaker detection where speech by only a single speaker is present. Three specific multi-speaker tasks have been defined: multi-speaker detection, speaker tracking, and speaker segmentation. Detection means determining whether or not a particular speaker is present in a speech segment. We use the term multi-speaker detection to distinguish this task from the conventional speaker detection task where only one speaker is present in the speech segment. Tracking means determining the intervals within a speech segment where a known speaker is speaking. Segmentation means determining the intervals within a segment that correspond to each of the unknown speakers present in the segment.

2 Multi-Speaker Detection

Multi-speaker detection is the task of determining whether a particular known speaker is present in a speech segment containing speech from multiple speakers. It may be viewed as an extension of the basic one-speaker detection task, and was added to the evaluations in 1999.

2.1 Evaluation Data

The primary data source for these evaluations has been the Switchboard Corpora ([3], see also [4]) of conversational telephone speech. While the basic one-speaker detection task uses data from a single conversational channel, these tasks use the summed channel signal. Thus the test segments each contain two conversing speakers. The two-speaker test segments were each selected to be approximately one minute in duration. The duration of speech by the individual speakers varied, with most in the 15-45 seconds range. For each test segment there were 22 trials, each with a single specified target speaker. Two of these trials were target trials, i.e., the specified target was one of the segment speakers. For each trial, systems were asked to provide both a true-or-false actual decision, and a numerical score indicating the likelihood that the target speaker was present in the test segment.

2.2 Comparison with One-Speaker Detection

Speaker detection with two speakers present is clearly a harder task than with just one speaker. This would be true even if the speech of the two speakers could be perfectly separated from each other. (See the discussion of the segmentation task below.) For suppose a one-speaker detection system has an operating point where the missed detection rate is p and the false alarm rate is q.

Consider a target trial for a two-speaker segment; i.e., the target speaker is one of the speakers in the test segment. The speech of the target in the test segment is correctly associated with the target with probability $1-p$, while that of the other segment speaker is incorrectly associated with the target with probability q. Assuming these decisions are independent of each other, the probability of a correct decision becomes

$$1 - p + q - (1 - p)*q = 1 - p + p*q$$

Thus the missed detection rate is

$$1 - (1 - p + p*q) = p * (1 - q)$$

Now consider a non-target trial; i.e., the target speaker is neither of the two segment speakers. For each segment speaker the false alarm probability is q, so together, with an independence assumption, the false alarm probability is

$$2q - q**2 = q * (2 - q)$$

Thus if p and q are both small (which is not entirely the case with current technology), then the false alarm rate is nearly doubled while the missed detection rate is little changed.

Figure 1 applies these ideas for one particular system from the 2001 evaluation. Results are presented for two-speaker trials where the test segment contained two speakers of the same sex, and for similar one-speaker trials where the test segments consisted of the single speaker portions of the two-speaker segments. The figure shows Detection Error Tradeoff (DET) performance curves (see [5]), based on the submitted likelihood scores, for the two-speaker task and for the one-speaker task, and shows the “predicted” two-speaker performance based on the above analysis. The performance deterioration for the two-speaker task is considerably larger

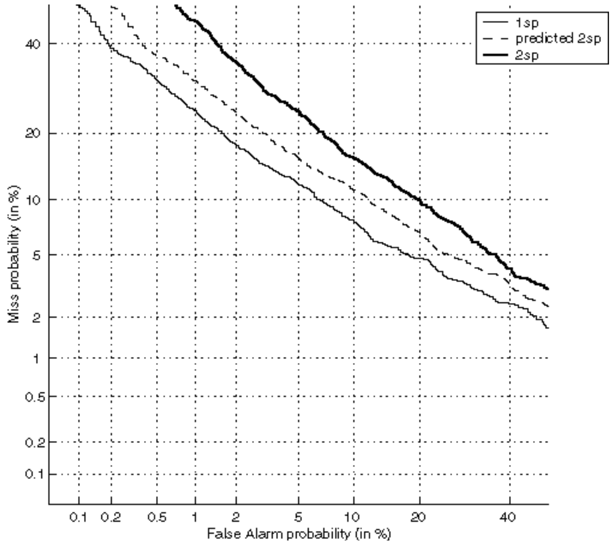

Figure 1: *DET Curves show performance for one system for one-speaker detection, "estimated" two-speaker detection, and actual two-speaker detection*.

than predicted, suggesting that speaker separation is a considerable part of the difficulty of this task.

2.3 Effect of Duration

The durations of speech by each segment speaker vary from close to zero to almost a minute or, alternatively, the percentage durations of each target speaker in test segments vary from close to 0 to close to 100. Figure 2 show one system's performance results when the target trials are partitioned according to the test segment durations of speech by the target speaker.

Figure 2 suggests that performance is enhanced up to the point where the target speech constitutes a clear majority of the total speech in the segment. Past results on one-speaker detection have suggested that performance gain with increasing duration

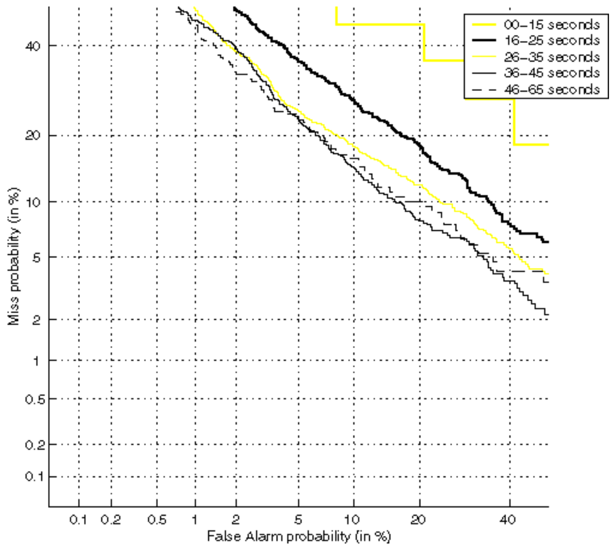

Figure 2: *DET Curves of two-speaker detection performance for one system as the true speaker speech duration in target trials*.

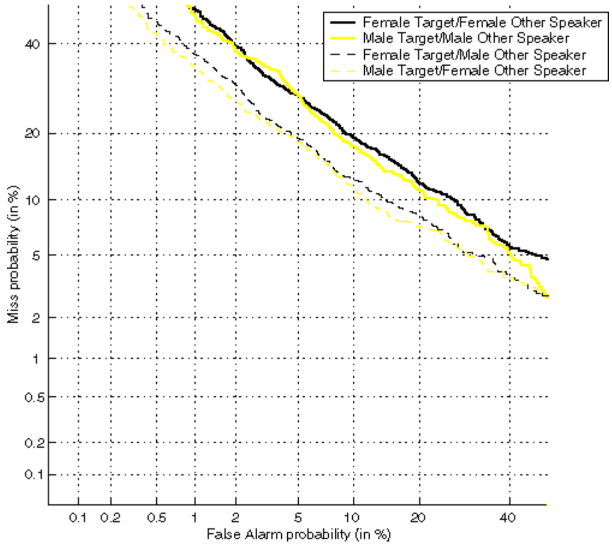

Figure 3: *DET Curves of two-speaker detection performance for one system as the gender mix varies for all trials.*

is limited once duration is at least 15 seconds. This is a further indication of the difficulty of the two-speaker task, and in particular the problem of separating the speech of the two speakers.

2.4 Gender Mix

The results presented up to this point have involved test segments where the two speakers were of the same sex, and this has indeed been part of the conditions of primary interest in these evaluations. The task should be less difficult when the speakers are of opposite sex, and this is illustrated for one system in Figure 3.

The four plotted curves cover the cases where both speakers are male, both are female, the target is male and the other female, and the target is female and the other male. The expected effect is seen, with the mixed gender segment having approximately half the false alarm rates of the same gender segments for fixed miss rates.

3 Speaker Tracking

When there is a speaker of interest, it may be desirable to determine not only whether the speaker appears in a multi-speaker segment, but to identify the specific intervals within the segment corresponding to the speaker. This is the speaker tracking task, which was added to the evaluations in 1999. A subset of the trials for the two-speaker detection task has been used for this task.

Tracking is scored as a detection task, measuring speech duration that is correctly or incorrectly classified as belonging to the known speaker. Speech intervals by the speaker may be missed; speech intervals by another speaker may be included, producing false alarms.

The objective of this task is to cluster the speech by speaker. It is not to determine exact speech boundaries, and it is recognized that speech detectors will differ slightly in the boundaries they determine. Therefore, in the 2000 evaluation, periods of non-speech, as determined by a NIST speech detector, were ignored for scoring purposes, as were "collar" periods of 250 msec. at the ends of all speech segments.

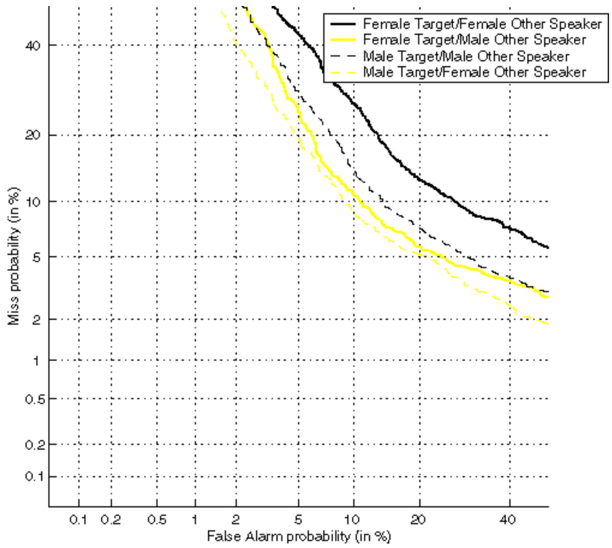

Figure 4: *Speaker tracking task performance DET Curves for one systems as gender mix of all test segments varies.*

Figure 4 shows tracking performance curves for one system as the gender mix of the test segments varies. The trials included here all involve test segments containing the target and one nontarget speaker. As expected, performance is enhanced when the other speaker is of opposite sex from the target. It is not clear why here it is the case of two female speakers that is most difficult. This was not the case for detection as shown in Figure 3.

The 2000 evaluation was the second to include the tracking task. The levels of performance shown for 2000 represent some improvement over those in 1999, but also suggest that this remains a difficult task and that performance may benefit considerably from further research.

4 Speaker Segmentation

A possible approach to the preceding tasks, used by at least some of the evaluation participants, is to cluster the speech into collections of intervals each believed to correspond to a single speaker. Then ordinary single-speaker detection may be applied to these to determine whether they contain the speaker of interest.

It was decided that the first part of this procedure might usefully be viewed as a separate task. The segmentation task thus requires associating the intervals found with specific, though unknown, speakers. It constitutes a type of clustering. This task was added to the evaluations in 2000.

The number of speakers present may be known or, perhaps more realistically, unknown. The 2000 and 2001 evaluations included a subtask where participants knew that exactly two speakers were present and a subtask where the number of speakers varied and was not specified.

The segmentation task, unlike the tasks discussed previously, is not a detection task, so results cannot be presented in terms of two error rates. A scoring metric was therefore specified which, in simple terms, may be expressed as follows: we consider all possible assignments of hypothesized speakers to different actual speakers, and take the one resulting in the smallest percentage of the speech (by duration) being assigned to an incorrect speaker. This percentage is then referred to as the segmentation error. As in the tracking task, non-speech intervals are ignored for scoring purposes, as are collar periods of 250 msec. at the ends of all speech intervals.

4.1 2-Speaker Subtask

This subtask used the same segments as the tracking task. These were known to contain speech by exactly two speakers.

For each test segment, the segmentation output could be transformed into a tracking result by associating each of the two clusters with a segment speaker. Let us suppose that the two clusters between them contain all of the speech in the segment and that the optimal association of clusters with speakers, as used to define segmentation error, is chosen. Then, for each of the two speakers, the segmentation error will equal the mean of the missed detection and false alarm tracking error rates for the speaker. It is therefore of interest to compare the segmentation error with the mean of the tracking missed detection and false alarm rates over the same segments, as presented for one system in Table 1.

Segment Speakers	Segmentation Error	Mean of Tracking Error Rates
Female	0.11	.16
Male	0.08	.12
Mixed	0.09	.10

Table 1: *Comparison for one system of segmentation error and tracking error (mean of missed detection and false alarm rates) on a common set of two-speaker test segments for male, female, and mixed gender segments.*

The mean tracking error rate is about fifty percent higher than the segmentation error rate for same sex segments, but only a little higher for mixed sex segments. This is perhaps due to the easier association of clusters with speakers when the genders differ. For segmentation as for the other tasks, female speakers are apparently more difficult than male. Surprisingly, the mixed gender case had a slightly higher error rate than the all male case.

4.2 n-Speaker Subtask

This subtask utilized data from the multi-language CallHome Corpus collected several years ago by the Linguistic Data Consortium ([6]). These conversations involved a call by a foreign visitor to the United States to a friend or family member in his or her home country. In many of these, there were more than two speakers participating, so summed two channel segments of up to five minutes duration, were used to obtain segments with as many as seven different speakers.

Figure 5 examines performance for one system as a function of the number of speakers present in the segment. The numbers of segments with each number of speakers is indicated. Also shown is what the segmentation error would be for a hypothetical system that associated all speech with a single speaker. The sizes of the white parts of the bars in Figure 5 thus show the system's average "value added" for segments with each number of speakers.

As might be expected, both the actual and hypothetical error rates increase with increasing numbers of segment speakers.

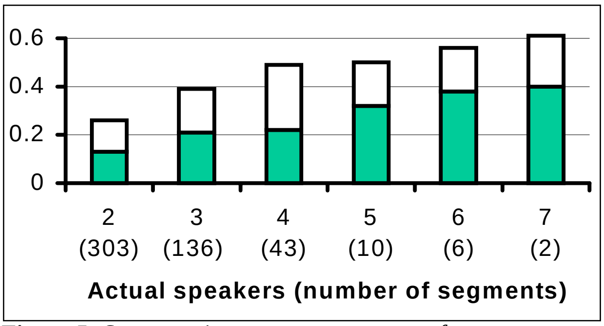

Figure 5: *Segmentation average error rates for one system as the number of segment speakers varies. Values in parentheses show the numbers of test segments used containing each number of speakers. The shaded lower part of each bar shows the actual segmentation error rate. The full bar shows what the error rate would be in a hypothetical system that associated all speech with a single speaker.*

Figure 6 examines the variation in error rate as a function of the six languages occurring in the test segments. A matrix of numbers of segments by language and number of speakers is also shown, as are the hypothetical error rates of a system that always finds a single speaker as discussed above.

The lowest average error rates occur with English and German segments, but this may be due to these languages having few segments with more than two speakers. Perhaps most notable is the apparent relative difficulty for the system of the Japanese segments, despite there being few of these with more than two speakers.

	Number of Speakers					
Language	2	3	4	5	6	7
Arabic	50	28	10	3	4	
English	49	7				
German	52	12	3			
Japanese	53	10	2	3		
Mandarin	47	50	18	2	1	
Spanish	52	29	10	2	1	2

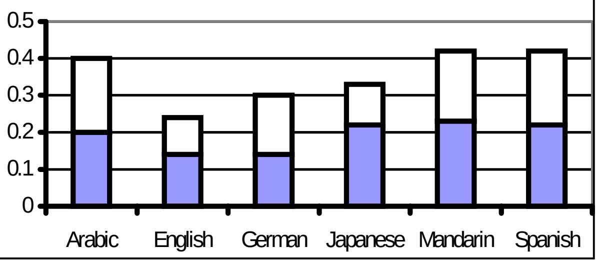

Figure 6: *Segmentation average error rates for one system as a function of the language used in the test segment. The full bars show the error rates for a hypothetical system assigning all speech to a single speaker. The matrix above shows the distribution of test segments by language and number of speakers.*

5 Conclusion

While one-speaker detection remains the central problem of speaker detection, there is significant interest in multi-speaker tasks. These tasks are clearly essentially harder than one-speaker detection. NIST intends to include these tasks in some form in future evaluations. It is hoped that increasing attention to them will lead to improved performance.

It is not easy to find appropriate data for these tasks that involve more than two speakers. There has been some recent interest, however, among speech recognition researchers, in collecting new corpora of meeting room speech data. This may be non- telephone data, but some conference calls may be included as examples of meeting data. Such data may be useful for speaker recognition as well as speech recognition. The authors would be interested in knowing of other appropriate data sources, particularly in languages other than English.

The NIST evaluations, it should be noted, are open to all interested participants. Recent evaluations have included participants from all over the world (see [7]). Information on these evaluations may be found at the website:

http:// ww.nist.gov/speech/tests/spk/index.htm

6 References

[1] NIST Speaker Recognition Evaluations – General Information, at *http:/www.nist.gov/speech/tests/spk/index.htm*

[2] Martin, A. and Przybocki, M., "The NIST 1999 Speaker Recognition Evaluation—An Overview", *Digital Signal Processing*, Vol. 10, Num. 1-3. January/April/July 2000

[3] Switchboard Corpora are available from the LDC at *http://www.ldc.upenn.edu/*

[4] Godfrey, J., et al., "SWITCHBOARD: Telephone Speech Corpus for Research and Development", *Proc. ICASS-92*, Vol. 1, pp. 517-520

[5] Martin, A., et al., "The DET curve in assessment of detection task performance". *Proc. EuroSpeeech* Vol. 4 (1997), 1895-1898

[6] CallHome Corpus is available from the LDC at *http://www.ldc.upenn.edu/*

[7] Martin, A. and Przybocki, M., "The NIST Speaker Recognition Evaluations: 1996-2000", *Proc. Odyssey Workshop*, Crete, June 2001