SPEAKER TRACKING AND DETECTION WITH MULTIPLE SPEAKERS

Kemal Sönmez [1] *Larry Heck*[2] *Mitchel Weintraub*[2]

[1]SRI International, Menlo Park, CA 94025
[2]Nuance Communications, Menlo Park, CA 94025
kemal@speech.sri.com {heck,mw}@nuance.com

ABSTRACT

We describe a speaker tracking and detection system, for Switchboard conversations, that uses a two-speaker and silence hidden Markov model (HMM) with a minimum state duration constraint and Gaussian mixture model (GMM) state distributions adapted from a single gender- and handset-independent imposter model distribution. Speaker tracking is used to segment speakers for detection, which is carried out by averaging frame scores of the Viterbi path and HNORM'ing via a novel parameter interpolation extension of HNORM for use with files of arbitrary lengths. Use of duration statistics augmenting the acoustic scores is also introduced via a nonlinear combination function. Results are reported on the NIST 1998 Multispeaker development evaluation dataset.

1. INTRODUCTION

As speech starts being exploited fully as an information source, multispeaker tracking and detection systems are increasingly in demand in a wide range of applications from indexing and archiving of broadcast news sources to software robot assistants that track dialogs and supply relevant information.

An early representative of the work on speaker detection in the presence of multiple talkers is the BBN Top-N 1-s classification algorithm in which the best N scoring 1-s segments are selected and used to compute the detection score. The Top-N's simplicity prevents it from addressing situations where the target speaker has less speech than the other speaker(s), or where two or more speakers share the utterance period evenly. More sophisticated approaches include BBN's subsequent approach in Siu [2] and Wilcox [3]. In [2], a single Gaussian mixture was used to represent speech (and another mixture was used to represent the noise). In [3], a single mixture model and a tied mixture model was used to represent the speakers. Both [2] and [3] focused on the problem of speaker segmentation without the use of training data of any speakers.

In this paper, we describe a speaker tracking and detection system for Switchboard conversations in the case where training data are available for the target speaker. The conversation is modeled as a two-speaker and silence hidden Markov model (HMM). A similar model was used earlier in [3]. In our model, Gaussian mixture model (GMM) state distributions are adapted from a single gender- and handset-independent imposter model distribution, and a minimum state duration is imposed. Speaker tracking is used to segment speakers for detection, which is carried out by averaging frame scores of the Viterbi path. For both tasks, handset effects are mitigated by HNORM [1] via a novel parameter interpolation extension of HNORM for use with waveform segments of arbitrary lengths. We also introduce a way to use duration statistics augmenting the acoustic scores via a nonlinear combination function. We test the system and report results on the NIST 1998 Summer development dataset.

2. SPEAKER CHANGE MODEL

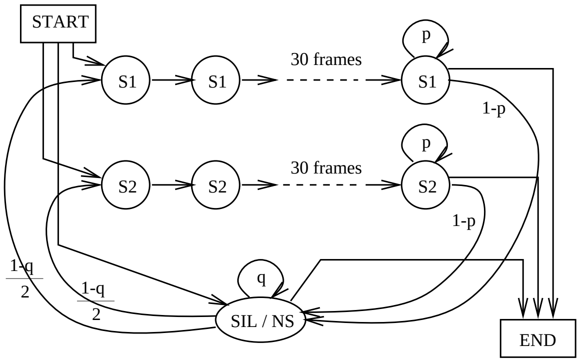

Figure 1. Speaker change model.

The model of the Switchboard conversation consists of an ergodic HMM (Figure 1) whose state distributions (GMMs with 512 G) are adapted from a single gender- and handset-independent imposter model distribution. The HMM consists of three states for modeling turns among talkers on channels A and B and silence. The silence model has the same structure as the imposter model. The talker states have a minimum duration of 0.3 s. For the speaker tracking and multiple speaker detection tasks, initially the same algorithm is run:

1. Likelihood scores: computation of target, imposter, and silence scores for each frame
2. Segmentation: Viterbi or forward-backward for posterior computation

For speaker tracking, once the waveform is segmented, likelihood ratios for each segment are computed from the

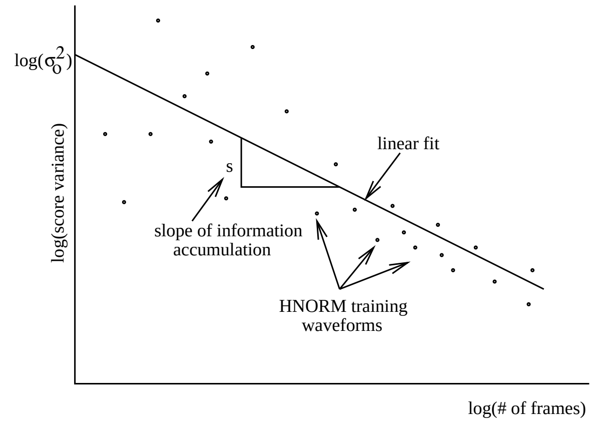

Figure 2. sHNORM

target and the imposter models. For multispeaker detection, average of scores from the frames segmented as target is augmented by statistics of duration to generate a score per test waveform. Single speaker detection is accomplished with the same GMMs and imposter model.

3. HNORM WITH VARIANCE MODELING

All scores are normalized with respect to handset variation via an extension of HNORM. In HNORM, the mean and the variance of scores of a speaker model on a set of imposter waveforms with the same handset are estimated, and then the scores are ZNORM'ed with the set of parameters fitting the handset type of the test waveform. For a given speaker model, we denote the scores on a set of imposter waveforms with handset α as $\{S_1^\alpha, S_2^\alpha, \ldots, S_K^\alpha\}$. Then, the first- and second-order statistics for the model on handset α are

$$\mu_\alpha = \frac{1}{K}\sum_{j=1}^{K} S_j^\alpha, \tag{1}$$

$$\sigma_\alpha^2 = \frac{1}{K}\sum_{j=1}^{K} (S_j^\alpha)^2 - \mu_\alpha^2. \tag{2}$$

HNORM normalizes the waveform i that has the handset type α as

$$\hat{S_i^\alpha} = \frac{S_i^\alpha - \mu_\alpha}{\sigma_\alpha}. \tag{3}$$

Note that Equations (1,2) assume that the K waveforms are of comparable size. The standard deviation estimates will vary significantly as a function of the number of frames in the waveform. Variance of the scores obtained from the GMMs would tend to

$$\sigma^2(N) = \frac{\sigma_0^2}{N} \tag{4}$$

if they were independent, where N is the number of frames used in scoring. Because of the inherent correlation in the speech signal, the information does not accumulate that fast. A more reasonable model for variance is

$$\sigma^2(N) = \sigma_0^2(\frac{1}{N})^s \tag{5}$$

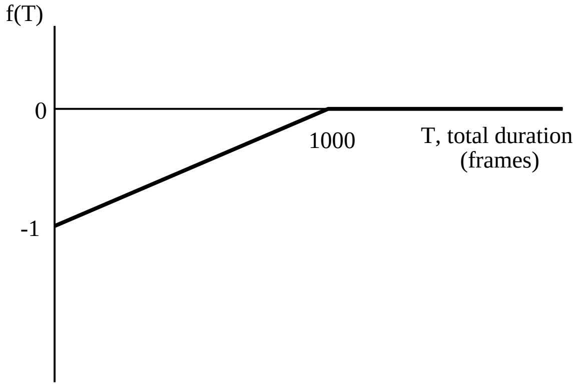

Figure 3. Duration score augmentation function

where $0 < s < 1$. Or

$$\log(\sigma^2(N)) = -s\log(N) + \log(\sigma_0^2). \tag{6}$$

This leads to a linear model in the log(σ^2)-log(N) domain with s as its slope (Figure 2), and log(σ_0^2) as its intercept. We estimate the two parameters from a set of scores obtained by running imposter waveforms of varying lengths against a given model. Once the s and log(σ_0^2) parameters for the $\sigma^2(N)$ function are estimated, each waveform/segment is normalized by the variance warranted by its duration:

$$\tilde{S_i^\alpha} = \frac{S_i^\alpha - \mu_\alpha}{\exp\frac{1}{2}\left(\log(\sigma_0^2) - s\log(N)\right)}. \tag{7}$$

In tracking, scores of segments labeled as belonging to a single speaker are sHNORM'ed, and in detection the average of all frame scores in all the labeled segments is sHNORM'ed according to Equation (7). sHNORM gives gains of 10-15% in various performance numbers in both tasks over no handset normalization.

4. SPEAKER DETECTION WITH DURATION AND ACOUSTIC SCORE COMBINATION

We propose a simple way to combine acoustic and duration information for multispeaker detection. The average of acoustic scores from the frames segmented as the target is augmented by statistics of duration via a thresholded nonlinear function to generate a score per test waveform. This is an *ad hoc* yet effective way to address the problem of the reliability of too few frames labeled as target speaker on which to average the scores. Scores averaged with less than a threshold size are decreased with a linear penalty. The parameters and the shape of the augmentation (penalty) function have been optimized empirically. Specifically, let S_a be the acoustic likelihood ratio (after sHNORM) score for the waveform, and let T be the number of frames for which the tracking algorithm has detected the target speaker. Then, the combined score is obtained by

$$S_c = S_a + f(T) \tag{8}$$

where $f(\cdot)$ is given by (Figure 3)

$$f(t) = (at + b)I_{[t<\tau]} \tag{9}$$

with $a = 0.001$, $b = -1$, and $\tau = 1000$. Since sHNORM normalizes the scores to the realization of a normal distribution with zero mean and unit variance, these parameters should be essentially independent of the specific type of acoustic scoring. Such a score combination has resulted in gains of 27% in equal error rate and detection cost function (DCF) in the multiple speaker detection task with respect to the sHNORM'ed acoustic scores.

5. NIST 1998 MULTISPEAKER DETECTION AND TRACKING TASK

In the NIST 1998 Multispeaker development evaluation [4], the two-speaker detection task is to determine whether a specified target talker is speaking during a given segment of conversational speech between two people, that is, a Switchboard call. The tracking task is to detect those time intervals (if any) during a given segment of speech when a specified target talker is speaking. An additional task, one-speaker detection task, is the same detection task on seperated Switchboard channels, that is,waveforms containing a single speaker.

The training in the Multispeaker development evaluation is the "two-session" condition where two separate waveforms of 1-min duration each are supplied as training for a single speaker. The test waveforms for the two-speaker detection and tracking tasks are 60 s long. The one-speaker task waveforms may vary between 0 and 60 s depending on the presence of the specified talker. There are 250 male and 250 female speakers with about 72,000 trials for the two-speaker detection task, 108,000 trials for the one-speaker detection task, and 4,000 trials for the tracking task.

6. RESULTS

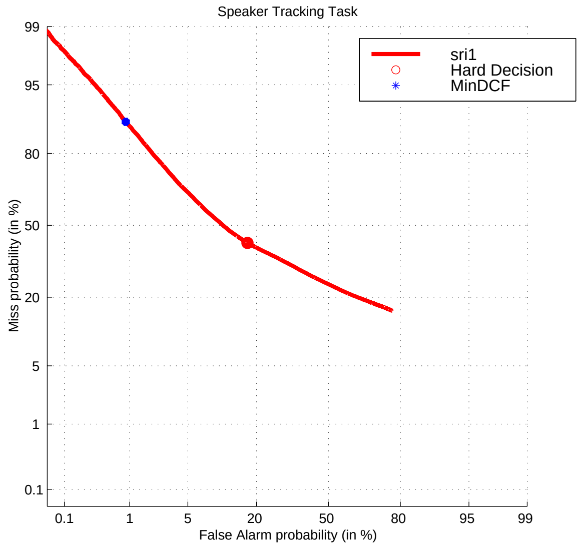

Figure 4. Tracking detection curve

The detection curve of the speaker tracking system is shown in Figure 4. The results of speaker detection on the NIST Evaluation are detailed in the detection curves in Figures 5 through 9. Figures 5 and 6 show the performance of the system without and with sHNORM, respectively, in the one-speaker and two-speaker testing conditions. It is observed that sHNORM helps significantly, 10-15% on average in various performance numbers. Figures 7 and 8 show the performance for waveforms with at least 25 s of data, eliminating cases where statistically there was very little material with which to work. The last figure compares the performance with that on the NIST 1998 speaker recognition evaluation with a similar system [4].

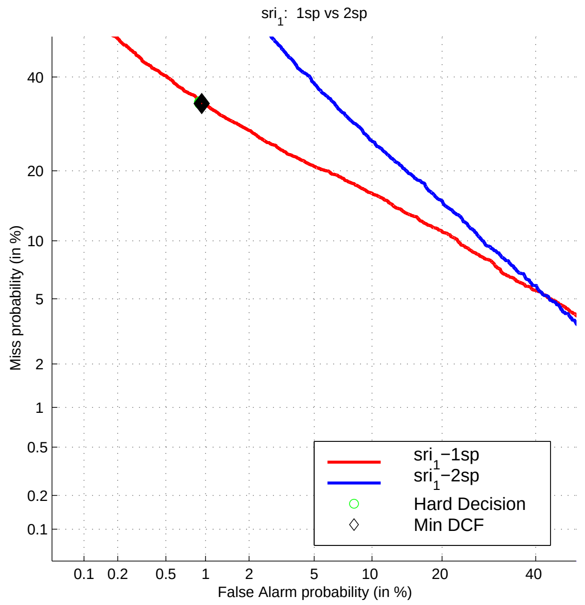

Figure 5. Detection curve: 1sp vs 2sp without sHNORM

7. SUMMARY

We have demonstrated the effectiveness of an HMM speaker change model for speaker tracking and multispeaker detection and introduced an extension of HNORM for waveform segments of varying lengths. We also propose a simple way to make use of duration statistics in multispeaker detection. The results have been presented on the NIST 1998 multispeaker development evaluation.

REFERENCES

[1] D. A. Reynolds. "The Effects of Handset Variability on Speaker Recognition Performance Experiments on the Switchboard Corpus. In *Proc. Int. Conf. on Acoust., Speech, and Signal Processing*, vol. I, pp. 113-117, Atlanta, GA, 1996.

[2] M.-H. Siu, G. Yu, and H. Gish, "An Unsupervised Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers", *Proc. of ICASSP*, vol. 2, pp. 189-192, San Francisco, 1992.

[3] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of Speech Using Speaker Identification", *Proc. of ICASSP*, vol. I, pp.161-164, 1994, Adelaide Australia.

[4] http://www.nist.gov/speech

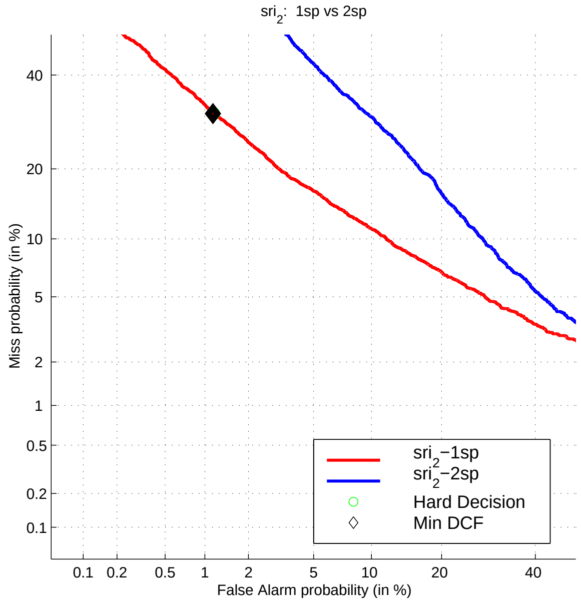

Figure 6. Detection curve: 1sp vs 2sp with sHNORM

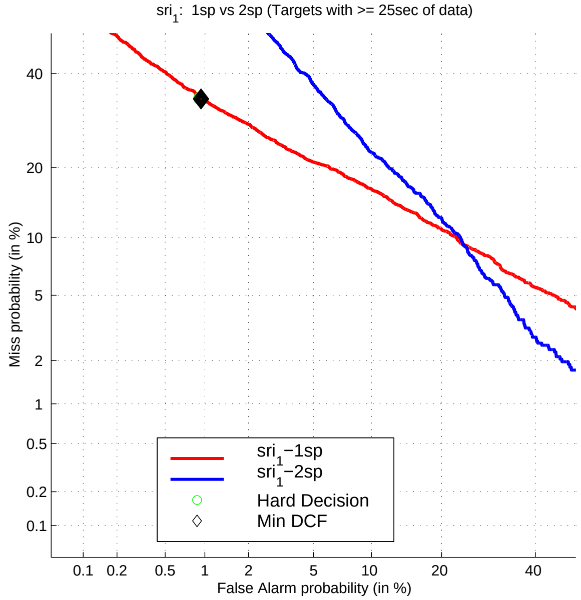

Figure 7. Detection curve: 1sp vs 2sp without sHNORM for a minimum duration of 30 s

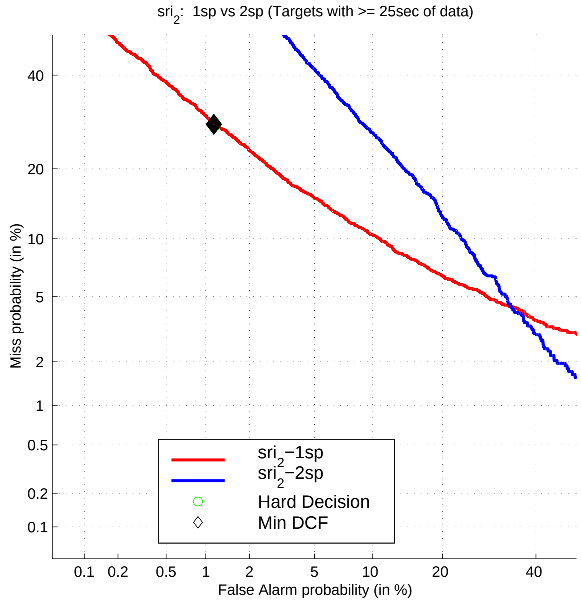

Figure 8. Detection curve: 1sp vs 2sp with sHNORM for a minimum duration of 30 s

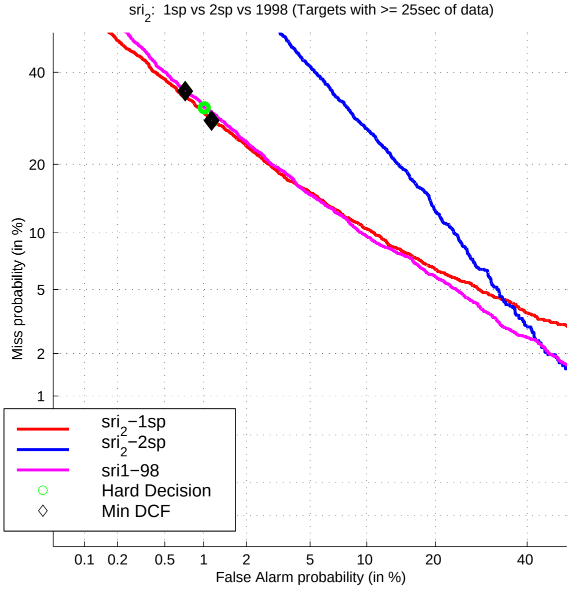

Figure 9. Detection curve: 1sp vs 2sp with sHNORM comparison with NIST 1998 Evaluation performance