Digital Signal Processing **10**, 133–142 (2000)
doi:10.1006/dspr.1999.0368, available online at http://www.idealibrary.com on IDEAL®

Multiple Speaker Tracking and Detection: Handset Normalization and Duration Scoring

Kemal Sönmez*, Larry Heck†, and Mitchel Weintraub†

*SRI International, Menlo Park, California 94025; and
†Nuance Communications, Menlo Park, California 94025

E-mail: kemal@speech.sri.com; heck@nuance.com, mw@nuance.com

Sönmez, Kemal, Heck, Larry, and Weintraub, Mitchel, Multiple Speaker Tracking and Detection: Handset Normalization and Duration Scoring, *Digital Signal Processing* **10** (2000), 133–142.

We describe SRI's speaker tracking and detection system in the NIST 1998 Speaker Detection and Tracking Development Evaluation. The system is designed for tracking switchboard conversations and 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 waveforms for speaker detection, which is carried out by averaging frame scores of the Viterbi path and normalizing for handset variation via a novel parameter interpolation extension of HNORM for use with waveform segments of arbitrary lengths. A short-duration penalty to augment the acoustic scores is also introduced via a nonlinear combination function. Results on the NIST 1998 Speaker Detection and Tracking Development Evaluation dataset are reported.

Key Words: speaker tracking; verification; handset normalization.

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. In this work, we report research on multispeaker tracking and detection by HMM-based speaker change models, handset normalization, and duration-based penalties. Most of the material and results in this paper have appeared in the conference article [3].

An early representative of the work on speaker detection in the presence of multiple talkers is the Top-N 1 second (1s) segments classification algorithm

1051-2004/00 $35.00

(developed by BBN) in which the best N scoring 1s segments are selected and used to compute the detection score. The Top-N approach is a simple method of computing statistics by filtering out the interfering speech and has proven to be effective in this capacity for verification of a single talker who dominates a multiple-talker utterance. However, its 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 [6] and Wilcox [7]. In [6], a single Gaussian mixture was used to represent speech (and another mixture was used to represent the noise). In [7], a single mixture model and a tied mixture model was used to represent the speakers. Both [6] and [7] focused on the problem of speaker segmentation without the use of training data for any speakers.

In this paper, our primary goal is to introduce a speaker tracking and detection system for two-channel telephone 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 [7]. 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. A second goal of this work is to develop extensions of handset normalization techniques for single speaker verification to speaker tracking and multiple speaker verification. For both tasks, handset effects are mitigated by a novel parameter-interpolation extension of HNORM [4] for use with waveform segments of arbitrary lengths. A final goal of the research is to introduce a way to use a short duration penalty to bias the acoustic scores via a nonlinear score combination function. We test the effectiveness of the system and normalization techniques and report results on the NIST 1998 Speaker Detection and Tracking Development Evaluation dataset.

We begin by introducing the task and the database in Section 2. The speaker change model is described in Section 3. Section 4 develops the extension of HNORM for use in speaker tracking and detection and reports its performance. Likewise, duration penalty to bias the scores is introduced together with experimental results in Section 5. We summarize our findings in Section 6.

2. TASK AND DATABASE

The 1999 NIST Multispeaker detection and tracking task has been detailed in this issue [2]. Basically, 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, is

the same detection task on separated switchboard channels, that is, waveforms containing a single speaker.

The Speaker Detection and Tracking Development Evaluation data had structure similar to that of the NIST 1999 Multispeaker Detection and Tracking Evaluation [2]. 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. These two waveforms have been recorded in two separate sessions from the same telephone number, and presumably with the same handset. The models trained (adapted) with these data are therefore most probably tuned to a single handset type, which makes handset normalization necessary, as explained in Section 4. The test waveforms for the two-speaker detection and tracking tasks are 1 min 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.

Performance measures for the task are the ROC and the detection cost function (DCF), which is basically the Bayes risk with preset miss and false alarm costs and target and imposter priors. A detailed description of DCF can be found in this issue [2].

3. SPEAKER CHANGE MODEL

Our model of the two-channel telephone conversation consists of an ergodic HMM (Fig. 1) with three states for modeling turns among talkers on channels A and B and silence. The state distributions are GMMs with 512 Gaussians. All the target speaker GMMs are adapted from a single gender- and handset-independent imposter GMM trained on Switchboard conversations from previ-

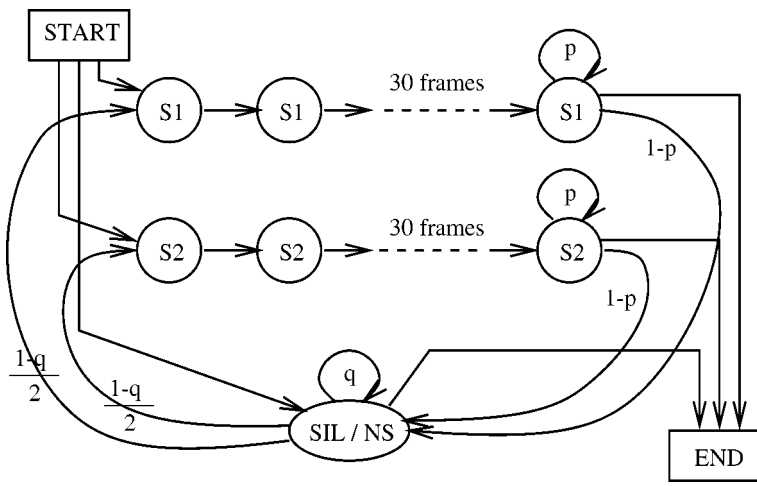

FIG. 1. Ergodic HMM speaker change model with minimum duration constraints.

ous NIST evaluations [8]. The details of adaptation for speaker verification are reviewed in this issue by its originators [1]. The silence/nonspeech model has the same structure as the imposter model and is trained on data segmented by an energy based speech detector. The talker states have a minimum duration of 0.3 s. This value has been observed by experiment to be effective for speaker tracking, an optimal trade-off between robust granularity (longer minimum durations) and noisy fine resolution (shorter or no minimum durations). The loop probabilities $p = q$ have been estimated on a development set as 0.999 and are fixed for all target speakers.

For the speaker tracking and multiple speaker detection tasks, initially the same HMM scoring algorithm is run:

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

The best path through the ergodic HMM automatically defines a segmentation by assigning a state for each frame. The posterior scores obtained via the forward–backward algorithm need to be thresholded to generate a hard decision for each frame to generate a segmentation. Experiments did not produce an edge for the more principled approach of computing the posteriors over the faster Viterbi computation; therefore the results reported are all with Viterbi segmentation.

For speaker tracking, once the waveform is segmented, likelihood ratios for each segment are computed from the 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. The performance of the speaker tracking system is tested on the development dataset and the resulting detection ROC curve is shown in Fig. 2.

4. HNORM WITH VARIANCE MODELING

Handset variation is the single greatest source of error in speaker verification over telephone channels on databases such as Switchboard. Effective handset normalization schemes against the vulnerability of current spectral matching techniques (features such as cepstrum) for speaker verification have been developed [4, 5]. In this section, we present a generalization of HNORM [4] useful for speaker tracking and detection with a wide range of waveform durations.

In the speaker tracking system, all scores are normalized with respect to handset variation via an extension of HNORM. HNORM is an extension of ZNORM, in which the mean and the variance of scores of a speaker model on a set of imposter waveforms with the same handset type are estimated, and then the scores are ZNORM'ed with the set of parameters fitting the handset type of

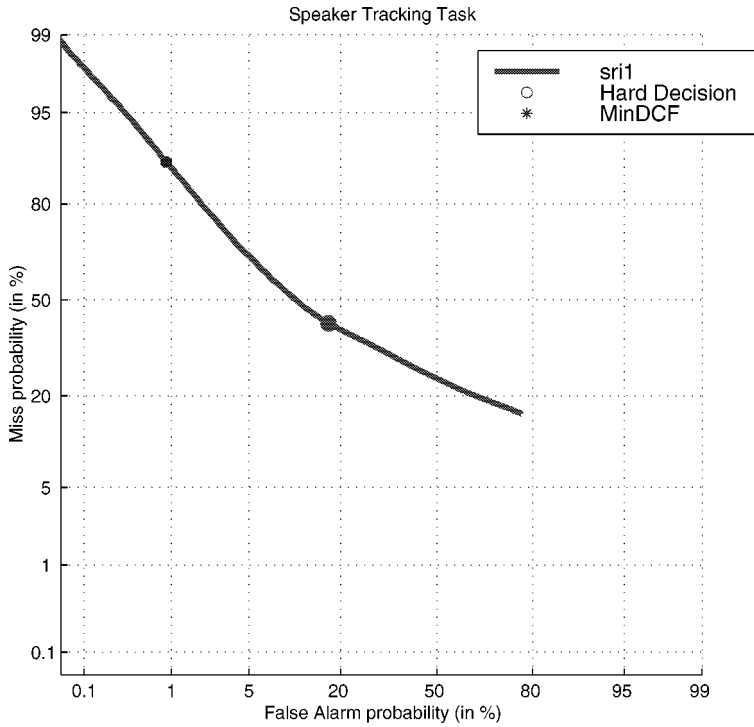

FIG. 2. Tracking detection curve.

the test waveform. For a given speaker model, we denote the scores on a set of imposter waveforms with handset type α as $\{S_1^\alpha, S_2^\alpha, \ldots, S_K^\alpha\}$. Then, the first- and second-order statistics for the model on handset type 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 score of 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 Eqs. (1), (2) assume that the K waveforms are of comparable size. The standard deviation of the scores will decrease significantly as a function of the number of frames in the waveform. Let $\{s_k\}_{k=1}^N$ denote the frame scores of a given waveform with a certain model. Assume for a moment that the frame scores s_i are i.i.d. with some distribution function F with mean μ_0 and variance σ_0^2. The overall score is the average of the frame scores:

$$S = \frac{s_1 + \cdots + s_n}{N}. \tag{4}$$

The mean of the random variable S is equal to that of the s_i, i.e., $\overline{S} = \mu_0$. The variance of S, however, is given by

$$\sigma_S^2 = \sigma_{(s_1+\cdots+s_n)/N}^2 = \frac{N\sigma_0^2}{N^2} \tag{5}$$

and is dependent on the number of frames, N. Therefore, if s_i were independent,

$$\sigma_S^2(N) = \frac{\sigma_0^2}{N}, \tag{6}$$

that is, the variance of the overall score would decrease in inverse proportion to N, 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

$$\sigma_S^2(N) = \sigma_0^2\left(\frac{1}{N}\right)^s, \tag{7}$$

where $0 < s < 1$, or

$$\log(\sigma_S^2(N)) = -s\log(N) + \log(\sigma_0^2). \tag{8}$$

In this model, correlation between the speech frames is taken into account to reduce the score variance more slowly than in the independent case. The slope of information accumulation, s, measures the rate at which variance is reduced. Once we have such a model, we can predict the variance with a given number of frames, and use the predicted variance in HNORM, resulting

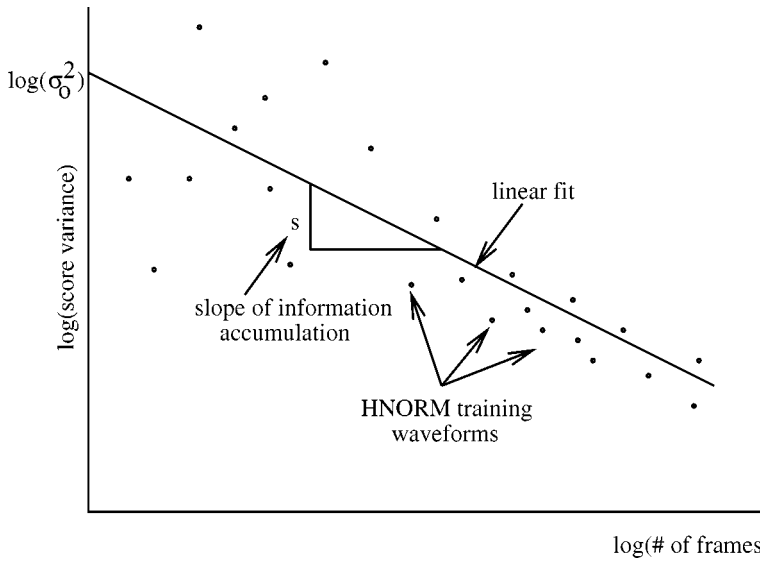

FIG. 3. sHNORM.

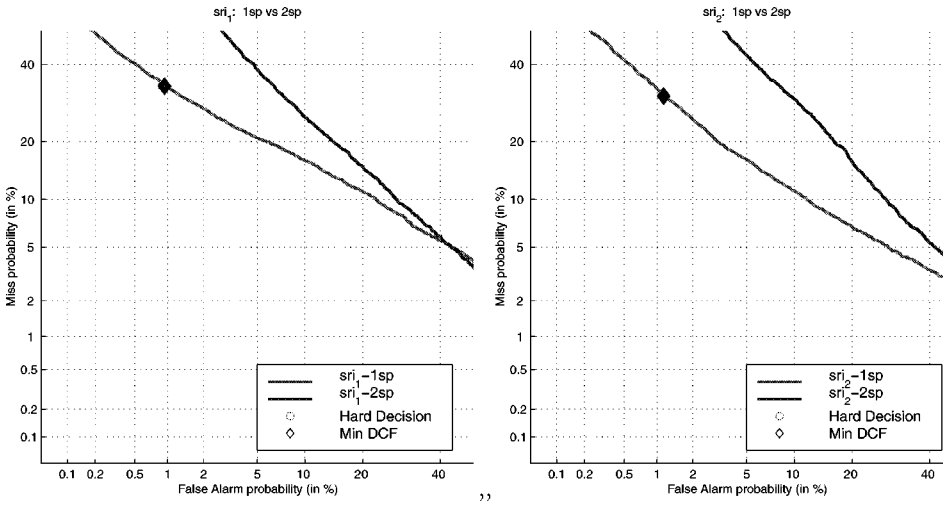

FIG. 4. Detection curve: 1sp vs 2sp without sHNORM (left) and with sHNORM (right).

in the introduction of sHNORM, an extension of HNORM with an estimated information accumulation slope.

The information accumulation model leads to a linear model in the $\log(\sigma_S^2)$–$\log(N)$ domain with s as its slope and $\log(\sigma_0^2)$ as its intercept (Fig. 3). 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 $\log(\sigma_0^2)$ parameters for the $\sigma_S^2(N)$ function are estimated by a simple linear line fit, each waveform/segment is normalized by the variance warranted by its duration:

$$\tilde{S}_i^{\alpha} = \frac{S_i^{\alpha} - \mu_{\alpha}}{\exp\frac{1}{2}(\log(\sigma_0^2) - s\log(N))}. \tag{9}$$

In tracking, scores of segments labeled as belonging to a single speaker are sHNORMed, and in detection the average of all frame scores in all the labeled segments is sHNORMed according to Eq. (9).

The detection cost function and equal error rate (EER) performance numbers of sHNORM are given in Tables 1 and 2 for the 1-speaker and 2-speaker detection tasks, respectively. It is observed that sHNORM gives gains of 5–15% in various performance numbers in both tasks over the baseline. Figure 4 shows the ROC performance of the system without and with sHNORM, respectively, in the one-speaker and two-speaker testing conditions.

TABLE 1

Performance on the 1-Speaker Task

System	DCF ($\times 10^3$)	EER
Baseline	57	14.6%
With sHNORM	50	13.4%

TABLE 2

Performance on the 2-Speaker Task

System	DCF ($\times 10^3$)	EER
Baseline	101	23.6%
With sHNORM	96	22.5%

5. SPEAKER DETECTION WITH DURATION AND ACOUSTIC SCORE COMBINATION

We propose a simple and *ad hoc* 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 a duration penalty 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 (1000 frames, determined experimentally) 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

$$S_c = S_a + f(T), \tag{10}$$

where $f(\cdot)$ is given by (Fig. 5)

$$f(t) = (at + b)I_{[t<\tau]} \tag{11}$$

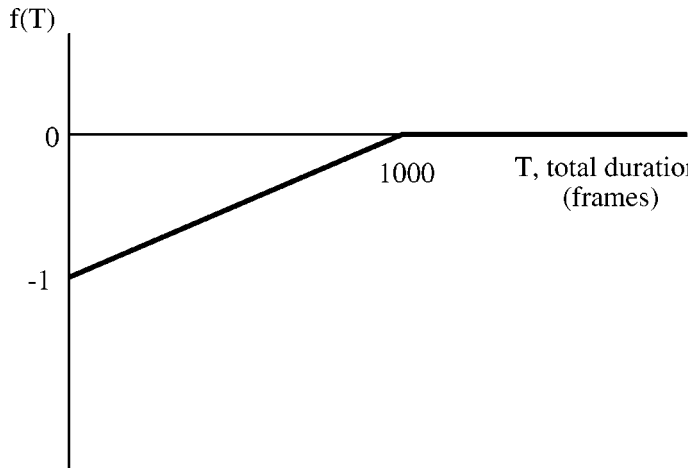

FIG. 5. Duration score augmentation function.

TABLE 3

Performance on the 1-Speaker Task

System	DCF ($\times 10^3$)	EER
Baseline	57	14.6%
With sHNORM	50	13.4%
With sHNORM and duration penalty	42	10.2%

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.

The DCF and the EER of the score combination with the duration penalty are given in Tables 3 and 4 for the 1-speaker and 2-speaker detection tasks, respectively. Duration penalties result in gains of up to 20% in EER and DCF in the multiple speaker detection task with respect to the sHNORMed acoustic scores.

6. SUMMARY

We have studied the effectiveness of an HMM speaker change model for speaker tracking and multispeaker detection. HMM framework allows the introduction of speaker continuity constraints by minimum state durations and speaker change penalties via the transition probabilities. The technique can be easily generalized to N talker conversations by introducing new speaker states in the ergodic HMM. We have also introduced an extension of HNORM for waveform segments of varying lengths. By modeling the information accumulation rate due to frame score correlation, sHNORM estimates a locus on the variance-number of frames plane rather than estimating a single variance. This is a principled alternative to using the same variance estimate for segments that vary by orders of magnitude in length or having several bins of lengths with common variances. We also propose a simple way to penalize short duration speakers in multispeaker detection. This technique relies on the randomness of the duration of the target speaker in waveforms and therefore is clearly dependent on the prior distribution of the durations. The results

TABLE 4

Performance on the 2-Speaker Task

System	DCF ($\times 10^3$)	EER
Baseline	101	23.6%
With sHNORM	96	22.5%
With sHNORM and duration penalty	80	18.3%

of both techniques on the NIST 1998 multispeaker development evaluation demonstrate significant improvement over the baseline HMM system.

REFERENCES

1. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., Speaker verification using adapted Gaussian mixture models, *Digital Signal Process.* **10** (2000), 19–41.
2. Martin, A. and Przybocki, M., The NIST 1999 speaker recognition evaluation—An overview, *Digital Signal Process.* **10** (2000), 1–18.
3. Sönmez, M. K., Heck, L. P., and Weintraub, M., Speaker tracking and detection with multiple speakers. In *Proc. Int. Conf. on Acoust., Speech, and Signal Processing, Budapest, Hungary,* 1999, Vol. V, pp. 2219–2222.
4. Reynolds, D. A., The effects of handset variability on speaker recognition performance experiments on the switchboard corpus. In *Proc. Int. Conf. on Acoust., Speech, and Signal Processing, Atlanta,* 1996, Vol. I, pp. 113–117.
5. Heck, L. P. and Weintraub, M., Handset-dependent background models for robust text-independent speaker recognition. In *Proc. ICASSP, Munich, Germany,* 1997.
6. Siu, M.-H., Yu, G., and Gish, H., An unsupervised sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In *Proc. of ICASSP, San Francisco,* 1992, Vol. 2, pp. 189–192.
7. Wilcox, L., Chen, F., Kimber, D., and Balasubramanian, V., Segmentation of speech using speaker identification. In *Proc. of ICASSP, Adelaide, Australia,* 1994, Vol. I, pp. 161–164.
8. Available at http://www.nist.gov/speech.