DETERMINING CO-LOCATION USING A SEQUENTIAL HYPOTHESIS TEST ON PATTERNS OF SILENCE

Wai-tian Tan *Ramin Samadani* *Bowon Lee* *Mary Baker*

Mobile and Immersive Experience Lab, Hewlett Packard Labs, Palo Alto, CA
{*wai-tian.tan, ramin.samadani, bowon.lee, mary.baker*}@*hp.com*

ABSTRACT

In everyday meetings, automatic association of *co-located* mobile devices would ease sharing of web-links, media, and other information. We propose a method that compares patterns of silence from device microphones to detect co-location of those devices. This method works with unsynchronized audio capture, requires only 100bps and preserves privacy. We show how to formulate pattern matching in a sequential hypothesis framework so that changes in co-location status (when people leave or join a meeting) can be determined promptly, and how to compute the likelihood ratio in practice. Using 16 hours of captured audio, we show that our approach can correctly determine device co-location with a low error rate of 0.05%, and can detect co-location changes 10 seconds faster than a similar decision rule based on a constant time window. Compared to a prior audio signature method, we achieve higher accuracy at 1/7 the bit rate.

Index Terms— sequential hypothesis test, voice activity detection, mobile device association.

1. INTRODUCTION AND PRIOR WORK

Sharing text, pictures and other digital information in meetings is part of our daily lives, yet the process is often tedious, requiring typing or emailing long URLs or codes. If we can reliably and continously determine which devices are co-located and are thus part of the same meeting, we can implement easier and more intuitve methods of content sharing, for instance through user interfaces in which drag-and-drop device icons automatically appear.

Zhang and Trott [1] argue for effectively determining device co-location by comparing device audio signatures. This is because people hearing the same conversation provides a human-centric concept of a meeting, while physical location alone, as determined for instance using WiFi-based methods, can inadvertently associate nearby people who are in different conference rooms. Device audio signatures also allow us to include the devices of people attending a meeting remotely, via teleconference, which WiFi-based methods do not. Nevertheless, a practical audio signature needs to be low bit-rate, preserve privacy by not revealing the content of a conversation, support accurate and prompt determination of co-location, and must be implementable in a distributed fashion.

The binary *silence signature* offers all the above advantages. It is obtained using a voice activity detector to classify each 10 ms block into 1 bit of silence (0) or voice (1). We send this low bit-rate (100 bps) silence signature to a server that compares signatures across devices and returns the identities of the co-located devices. Preserving privacy, the silence signature does not reveal the conversation.

One alternative to matching silence signatures is to correlate either the audio signals or the windowed energy of those signals. We find this works poorly since people are at different locations in the room. The resulting mixed signal has audio components from each person but with different loudness at each device. Our approach using silence signatures works better because it is robust to these loudness differences.

The most similar prior work [1], proposes an audio co-location signature based on quantized phase. There are two essential advantages of a silence signature over quantized phase. First, compression can destroy phase information [1], meaning phase-based methods cannot generally extend to remote participants. Second, phase is inherently sensitive to alignment errors when the frames used to compute phase are not aligned temporally across devices. We show in the results section that this can cause significant drop in performance, and that a silence signature is more accurate overall. Furthermore, a quantized phase signature contains more information than a silence signature, requiring seven times higher bit rate.

Audio signatures have been extensively used in music search [2]. One key difference between co-location and music search is the access to a clean reference signal for music search [1]. In contrast, audio signatures to determine device co-location are all computed from noisy, distorted signals.

A sound-emitting method [3] uses sentences automatically generated from public keys and *vocalized* by a text-to-speech system to establish secure pairing between devices with the aid of humans for manual authentication. In contrast, our method uses the unmodified acoustic environment and does not require explicit user interaction.

There are non-audio ways of associating devices [4] such as bump [5]. These methods generally do not extend to re-

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

mote participants of tele-conferences, because they require physical *bumping* together of the devices.

Our contribution in this paper is three-fold. First, we propose use of a silence signature for determining co-location and evaluate its effectiveness. Second, we develop practical approximations of the joint pdf for the two silence signatures we compare. We use statistical models of the transitions between speech and silence from the two signals to allow computation of the likelihood ratio for hypothesis testing. Third, we apply sequential hypothesis testing [6, 7] to adaptively choose the smallest possible window size to reduce decision latency when co-location status changes (when someone joins or leaves the meeting). We compare the performance of our method with a constant window method, and a related prior method [1].

2. FORMULATION

Given current silence signatures $\mathbf{s}^{(0)}$, $\mathbf{s}^{(1)}$ from two devices, we want to decide between these two hypothesis

θ_0: $\mathbf{s}^{(0)}$ and $\mathbf{s}^{(1)}$ are signals from different locations

θ_1: $\mathbf{s}^{(0)}$ and $\mathbf{s}^{(1)}$ are co-located signals

by testing observations $\mathbf{x} = [\mathbf{s}^{(0)} \;\; \mathbf{s}^{(1)}]^T$ over a causal window τ. A straightforward approach uses constant τ independent of the content of $\mathbf{s}^{(0)}$ and $\mathbf{s}^{(1)}$. However, a large τ entails long decision latency when co-location state changes, whereas a small τ is susceptible to higher mis-classification. Instead, we seek a sequential solution where progressively larger observation windows of $\tau_1 \leq \tau_2 \leq \tau_3 \leq \ldots$ are tried in sequence until we are confident about our decision. In a sequential analysis framework [6, 7], this is the same as progressively testing

$$\frac{\beta}{1-\alpha} < \frac{f(\mathbf{x}(\tau_i)|\theta_1)}{f(\mathbf{x}(\tau_i)|\theta_0)} < \frac{1-\beta}{\alpha} \qquad (1)$$

until the middle likelihood ratio term falls below $\frac{\beta}{1-\alpha}$ when we declare θ_0 is true, or rises above $\frac{1-\beta}{\alpha}$ when we declare θ_1 is true. The quantities α and β are the desired mis-classification rate when θ_0 and θ_1 are true, respectively.

Taking the logarithm of (1), and choosing $\alpha = \beta$, we can simplify the decision rule as

θ_0 if $LLR(\tau_i) < -\theta$

θ_1 if $LLR(\tau_i) > \theta$

where LLR is the log likelihood ratio, and $\theta = \log \frac{1-\beta}{\alpha}$ is the decision threshold. We try τ_{i+1} if $|LLR(\tau_i)| \leq \theta$.

2.1. Likelihood under different locations: $f(\mathbf{x}(\tau)|\theta_0)$

We can model durations of speech and silence in conversation by exponential and shifted exponential ($p(t) = e^{-(t-t_0)}$ for $t \geq t_0$) distributions, respectively [8, 9], where t_0 is the minium silence duration produced by a voice activity detector, and is set to 0.1s in this paper. Assuming independence of individual silence and speech durations [9], we can express $f(\mathbf{x}(\tau_i)|\theta_0)$ as products of the probability of each silence and speech duration over time window τ_i across both signals. We compute the parameters for the exponential and shifted-exponential distributions empirically from $\mathbf{s}^{(0)}$ and $\mathbf{s}^{(1)}$ as the reciprocal of the average voice and silence durations, respectively.

2.2. Likelihood under co-location: $f(\mathbf{x}(\tau)|\theta_1)$

The pdf of joint observation $\mathbf{x}(\tau_i)$, given by $f(\mathbf{x}(\tau_i)) = f(\mathbf{s}^{(0)}(\tau_i), \mathbf{s}^{(1)}(\tau_i))$, can be written as

$$f(\mathbf{x}(\tau_i)) = f(\mathbf{s}^{(1)}(\tau_i)|\mathbf{s}^{(0)}(\tau_i)) \cdot f(\mathbf{s}^{(0)}(\tau_i)), \qquad (2)$$

where we can readily compute $f(\mathbf{s}^{(0)}(\tau_i))$ from the statistics described in Section 2.1. We approximate the conditional term $f(\mathbf{s}^{(1)}|\mathbf{s}^{(0)})$ as follows. We first find all begin times $b_j^{(0)}$ and end times $e_j^{(0)}$ of each silence period in $\mathbf{s}^{(0)}$. We then find the corresponding (i.e., closest) begin times $b^{(1)}(b_j^{(0)})$ and end times $e^{(1)}(e_j^{(0)})$ in $\mathbf{s}^{(1)}$. We then approximate the conditional probability as

$$\hat{f}(\mathbf{s}^{(1)}|\mathbf{s}^{(0)}) = \prod p_b(b_j^{(0)} - b^{(1)}(b_j^{(0)}))p_e(e_j^{(0)} - e^{(1)}(e_j^{(0)}))$$

where $p_b(d)$ and $p_e(d)$ are probabilities that co-located signatures have silence begin and end times that deviate by offset d, respectively, and are determined *a priori* via training.

We derive p_b and p_e from 7 co-located recordings from various devices in an actual project meeting that lasted one hour. We collected the silence duration as well as the offsets for all silence begin and end times. We find that the statistics of p_b and p_e are similar and not worthy of separate accounting. The distributions of offsets for silence periods shorter than 0.3 seconds and longer than 0.3 seconds are shown in the left and right plots of Fig. 1, respectively. Their similar shapes and symmetry suggest that it is not necessary to condition on silence duration, and it suffices to consider offset magnitude.

With these simplifications, the resulting empirical pdf of offset magnitude is shown in gray in Fig. 2. This is further

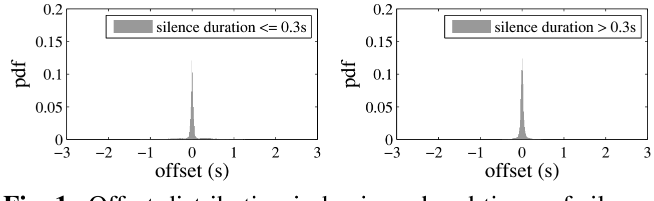

Fig. 1. Offset distribution in begin and end times of silence period of different durations for co-located silence signatures.

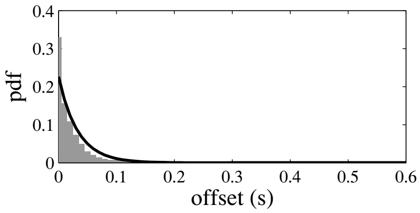

Fig. 2. Offset magnitude in silence begin and end times for co-located silence signature, with empirical distribution in gray and model in black.

approximated by the black curve in Fig. 2 by using an exponential distribution for offsets less than 0.3 seconds, and a uniform distribution for offsets between 0.3 to 3 seconds. Offsets larger than 3 seconds are clipped to 3 seconds. Both $p_b(d)$ and $p_e(d)$ are computed by looking up $|d|$ against the black curve.

With $\hat{f}(\mathbf{s}^{(1)}|\mathbf{s}^{(0)})$, we can compute $f(\mathbf{x})$. Rather than directly using (2), we instead use

$$f(\mathbf{x}) = \frac{1}{2}[f(\mathbf{s}^{(1)}|\mathbf{s}^{(0)}) \cdot f(\mathbf{s}^{(0)}) + f(\mathbf{s}^{(0)}|\mathbf{s}^{(1)}) \cdot f(\mathbf{s}^{(1)})] \quad (3)$$

to ensure symmetry in the arguments $\mathbf{s}^{(0)}(\tau_i)$ and $\mathbf{s}^{(1)}(\tau_i)$.

3. RESULTS

This section shows results 1) comparing accuracy of sequential versus constant windows; 2) comparing accuracy of sequential versus the method in Zhang [1]; 3) comparing decision latency for the sequential versus constant window method, including the case of a remote participant in a teleconference.

We first compare the accuracy of using an adaptive window as determined by the sequential method versus that of a constant window and that of the quantized phase method [1]. Our results use 5 audio recordings of 1 hour each. Three recordings, A, B, C, are from a different work meeting than that used to derive statistics in Section 2, while recordings D and E are from other unrelated meetings. For adaptive windows, we employ the sequential hypothesis testing starting with an intial window of τ_1 = 3 seconds, and increase in 1 second steps until the decision threshold θ is crossed. Co-location of A with B to E is evaluated every second. Fig. 3 shows the results expressed in receiver operating characteristics (ROC). With $\theta = 10$, the average window size is 4.3 seconds, and we already achieve a true positive rate (TPR) of 0.97, with a false positive rate (FPR) of 0.28%. Using a constant window of 5 seconds results in significantly worse TPR of 0.95 and FPR of 2.1%. Moving to $\theta = 30$, with an average window size of 10.8 seconds, the adaptive windows method achieves TPR of 0.997 and a perfect FPR. In contrast, using a constant window of size 11 seconds would result in worse TPR of 0.978 and FPR of 0.1%. Further testing with 16 hours of recordings including more diverse settings such as a noisy cafeteria has shown that we can achieve a TPR of 0.9995 (0.05% error) and a perfect FPR of 0. We do not trace a curve as we vary θ since the trade-off is not between TPR and FPR. Instead, larger θ corresponds to improvement in *both* TPR and FPR at the expense of a larger average window size.

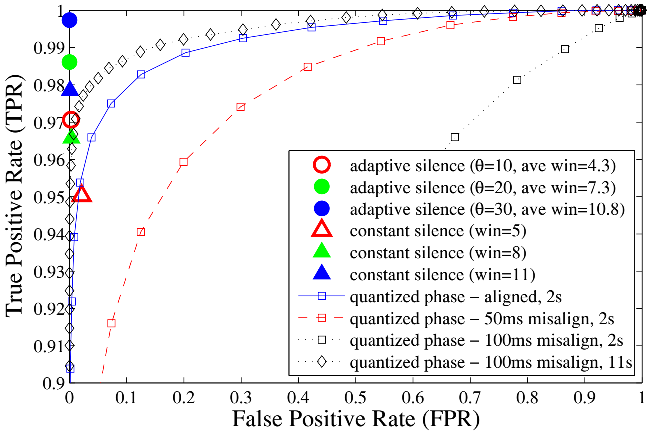

Fig. 3. Receiver operating characteristics for various approaches zoomed in for $0.9 \leq \text{TPR} \leq 1$.

The quantized phase method [1] computes quantized phase for 0.4s frames that are overlapping by 0.2s. Co-location is determined by comparing quantized phase in 2-second windows. With frames overlapping by 0.2s, a worst-case misalignment between two devices that determines frame boundary independently is 0.1s, with an average misalignment of 0.05s. From Fig. 3, we see the performance of quantized phase using the default 2-second windows lags that of our silence signature significantly, even with frame alignment. The performance gap increases drastically as frame misalignment increases. Nevertheless, the performance of quantized phase increases significantly when the window size increases to 11 seconds, achieving near perfect TPR and FPR with aligned frames. When alignment cannot be guaranteed, we see from Fig. 3 that silence signature provides significantly superior TPR and FPR over quantized frames with 11s windows and 0.1s misalignment while consuming only 100 bits per second compared to 700 for quantized phase.

Figs 4 and 5 shows the time trace of computed log likelihood for adaptive window with $\theta = 30$ and constant window of 11 seconds, respectively. We see that using adaptive windows is much more successful in separating the true positives (black) from true negatives (gray) than using constant windows. The corresponding trace for quantized phase using 11 seconds window and 0.1s frame misalignment is shown in Fig. 6. Again, silence signature with adaptive window provides superior separation.

We next illustrate the reduction in decision latency when co-location status changes. In this experiment, Ramin and Mary are in a conference room talking with Dan, who is re-

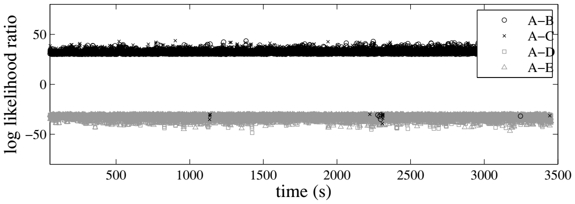

Fig. 4. Time trace of log likelihood ratio using adaptive window with $\theta = 30$.

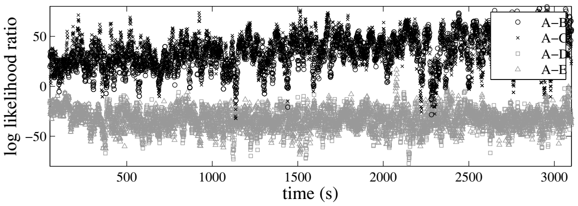

Fig. 5. Time trace of log likelihood ratio for constant window of 11s.

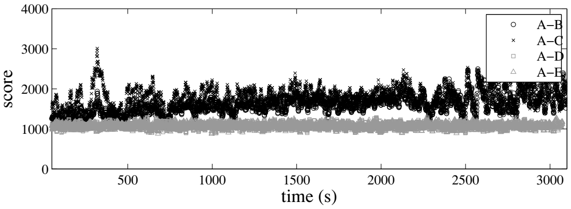

Fig. 6. Time trace of score computed using quantized phase with misalignment of 0.1s using 11s window.

mote (attending by telephone). At time 105 seconds, Mary leaves the room and returns at time 310 seconds. The adaptive window sizes are shown in Fig. 7 between Dan-Ramin and Mary-Ramin. Due to more distortion caused by the phone system, we need a larger average window size of 16 seconds between Dan-Ramin versus that of 11 seconds between Mary-Ramin.

Fig. 8 shows the log likelihood ratio evaluated every second using adaptive windows. Mary is determined to have left the room at time 116 seconds, re-entering at time 320 seconds, for a decision latency of 11 and 10 seconds, respectively. Fig. 9 shows the corresponding result employing a constant window size of 14 seconds – the average window size of Fig. 8. With constant window size, the maximum likelihood decision is to choose θ_0 or θ_1 depending on whether LLR is negative or positve, respectively. We see that the small window causes four mis-classifications between Dan-Ramin, at times 79, 231, 380, and 449 seconds. One remedy to reduce such mis-classification is to use a larger constant window. We have determned empirically that a window size of 50 seconds is necessary to prevent mis-classification between Dan-Ramin. Corresponding results for a constant window of 50 seconds are shown in Fig. 10. Nevertheless, the times in which Mary is determined to have left and re-entered the room are 127 and 340 seconds, respectively. This represents an additional 11 and 20 seconds worth of delay compared to the use of adaptive windows.

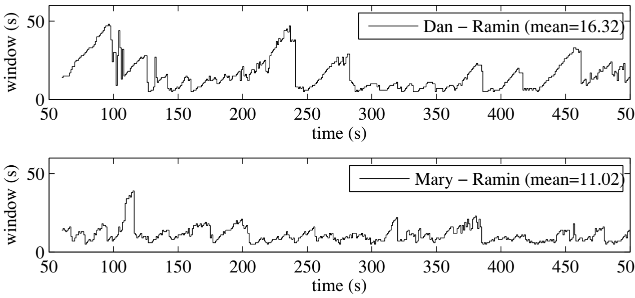

Fig. 7. Window size from sequential hypothesis test.

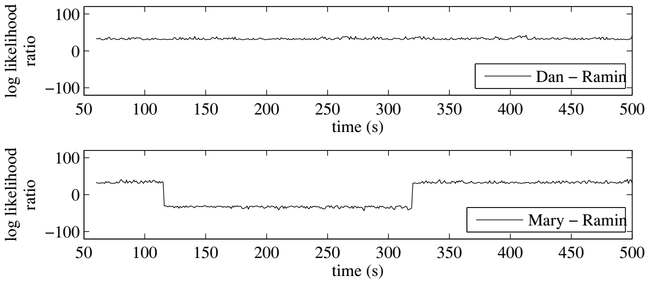

Fig. 8. Log-likelihood ratio using adaptive window.

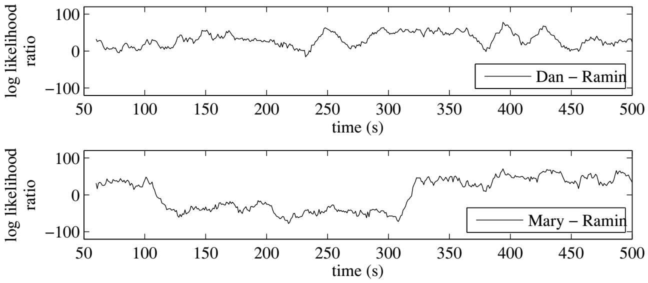

Fig. 9. Log-likelihood ratio using constant window of 14 seconds, the average window size of Fig. 8.

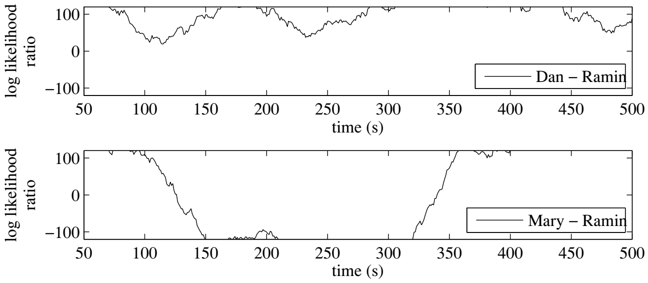

Fig. 10. Log-likelihood ratio using a constant window of 50s.

4. CONCLUSIONS

In this paper, we propose the use of silence signatures for determining device co-location, describe how to formulate this in a sequential hypothesis testing framework, and show that we achieve superior accuracy compared to a non-sequential test and a prior method based on audio signatures [1].

5. REFERENCES

[1] B. Zhang and M.D. Trott, "Reference-free audio matching for rendezvous," in *Proc. ICASSP*, March 2010, pp. 3570–3573.

[2] V. Chandrasekhar, M. Sharifi, and D.A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications," in *Proc. 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[3] M.T. Goodrich, M. Sirivianos, J. Solis, C. Soriente, G. Tsudik, and E. Uzun, "Using audio in secure device pairing," *International Journal of Security and Networks*, vol. 4, no. 1, pp. 57–68, 2009.

[4] M. Chong and H. Gellersen, "Classification of spontaneous device association from a usability perspective," in *Proc. The 2nd Int. Workshop on Security and Privacy in Spontaneous Interaction and Mobile Device Use (IWSSI/SPMU)*, March 2010.

[5] "Bump," http://bu.mp/company, November 2012.

[6] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, pp. 117–186, 1945.

[7] G.B. Wetherill, *Sequential methods in statistics*, Chapman and Hall, second edition, 1979.

[8] P. Brady, "A model for generating on-off speech patterns in two-way conversation," *Bell System Technical Journal*, vol. 48, pp. 2445–2472, 1969.

[9] ITU-T P.59, "Artificial conversational speech," 1993.