Online Detection of Speaking from Respiratory Measurement Collected in the Natural Environment

Md. Mahbubur Rahman[⋆], Amin Ahsan Ali[⋆], Andrew Raij[†], Emre Ertin[⋄]
Mustafa al'Absi[‡], Santosh Kumar[⋆]

University of Memphis[⋆], University of South Florida[†], The Ohio State University[⋄],
University of Minnesota Medical School[‡]

ABSTRACT

We present a novel method to detect when a person is speaking using respiratory measurements collected in the natural enviroment. A speaker's respiration pattern is sampled from a respiratory inductive plethysmograph (RIP) band worn around the speaker's chest. Ratio of inhalation duration to exhalation duration (IE ratio) has traditionally been used to detect speaking in controlled lab environment [7]. But we find that IE ratio is inadequate in the natural environment. We propose several new features to be used along with IE ratio. Using various statistics over these features in a decision tree, we are able to obtain >93% accuracy in classifying respiration measurements into "speaking" or "silence" states with 10-fold cross validation. Our demonstration will show real-time capture of the respiration signal, computation of features, and detection of speaking and silence all on a mobile smartphone.

1. INTRODUCTION

Reliable, low-power, unobtrusive detection of speaking events in natural environments has several uses. First, it can be used in scientific studies of conversation (or social interaction). Second, automated unobtrusive speaking detector can be used by people with social anxiety to monitor and improve their social interaction. Third, it can enable context awareness in smartphones; for example, a conversation-aware smartphone can automatically divert incoming calls to voice mail [3] during ongoing conversations.

State-of-the-art methods for detecting speaking state in natural environments primarily use audio signals from microphones of mobile phones. They extract frequency domain features, such as spectral flux, spectral centroid, relative spectral entropy, bandwidth, etc from the audio signal to detect voice [6, 9]. However, this approach has several disadvantages. First, use of microphones for recording audio raises privacy concerns. A privacy study in [5] shows that 91.3% of the participants are adamantly not willing to be audio recorded since they were feeling uncomfortable. 75% of them remained uncomfortable even if audio is recorded only in the frequency domain. Moreover, users of audio recording devices (e.g. Personal Audio Loop) are also greatly concerned about the privacy of others (conversation partners, passers by) whose data might be captured by the device [4]. They raised the issue that recording audio without consent from others around him/her may cause problems in their social relationships. Second, microphone-based speaking detection is not speaker-specific because the microphone can pick up the voice of anyone nearby. Thus, additional signal processing is needed to detect the speaking of a specific individual using a microphone. Third, mobile phones have computational and energy limitations. Frequent microphone sampling and feature computation in the frequency domain has a high computational and energy cost. Lastly, microphone occlusion can make it difficult to collect audio signals of sufficient quality for speaking detection. Microphone occlusion is common in natural environments because people tend to keep mobile phones in pockets and purses.

Using respiration measurements to detect speaking addresses all of the above shortcomings. But it requires wearing an unobtrusive respiration band around speaker's chest. Respiration band is also used in stress inferencing [8], physical activity detection [2], exposure to pollutants. If respiration band is used in any scientific study, speaking detection from the respiratory measurements does not increase any extra burden to the subject.

In this demonstration, we will show live detection of speaking state on a mobile phone that wirelessly captures respiration measurements. We will demonstrate a set of respiration features that distinguish speaking from silence, and show the real-time classification of these features into speaking and silence on a mobile smartphone.

2. OUR APPROACH

We implemented our system using Android SDK [1] and test it using Android G1 smartphone. We compute a set of features from the respiration signal, as described below, to train a decision tree classifier(J48). We select decision tree since it provides high accuracy at a low computational cost compared to other machine learning algorithm. The decision tree receives features computed from the raw signals captured wirelessly on the phone and produces the infereces.

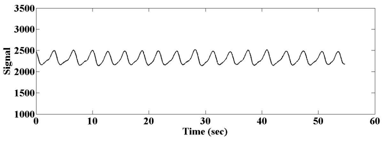

Figure 1: Respiration signal during silence.

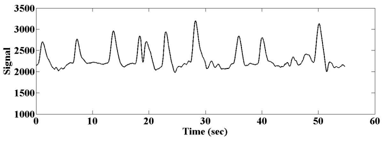

Figure 2: Respiration signal during speaking

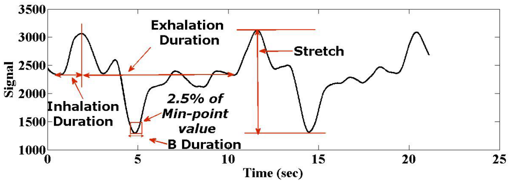

Figure 3: Illustration of different features

2.1 Respiration Signal

Figures 1 and 2 depict respiration signals that corresponds to silence and speaking events. The vertical axis refers to the ADC output from the respiration sensor (sampled at 64Hz) which corresponds to the lung volume. All the features used for the classification are based on the time-domain representation of the signal.

2.2 Feature Computation

Computation of the features involves the identification of respiration cycle which is composed of an inhalation and an exhalation period. Thus a respiration cycle starts from a valley that corresponds to the start of an inhalation and ends at another valley that marks the start of the next inhalation. Inhalation duration corresponds to the time elapsed from a valley of a signal to the next peak, which denotes the maximum expansion of the chest in the respiration cycle (see Figure 3) and exhalation duration corresponds to the time duration between the peak and the next valley. Exhalation duration during speech tend to be longer than that of silence. **IE ratio** is defined as the ratio of inhalation to the exhalation duration of a respiration cycle.

IE ratio has traditionally been the main feature for detecting speaking [7]. Use of only IE ratio did not prove to be effective enough in natural enviroment. The classification accuracy is observed to be approximately 78%. We propose several new features from the respiration signal. Using various statistics over these new features along with IE ratio, we achieve >93% accuracy.

New Features. We propose three novel features. **First Difference of Exhalation** is derived by taking the first order differences of the exhalation durations. This difference is observed to be lower in silence than that of speaking events. **Stretch**, the difference between the amplitude of the peak and the minimum amplitude the signal attains within a respiration cycle (see Figure 3), is found to be large during speaking events. Lastly, **B-Duration** defined as the time the signal continues to stays within 2.5% of the minimum amplitude, is also found to be longer during speaking events.

Feature Statistics. Although the features are defined for each respiration cycle, to mitigate the effect of noise and outliers (e.g. spikes), we compute statistics over a 30 sec window of respiration signal. Mean, median, standard deviation and the 80th percentile value are found to be useful features for classification. Thus, a total of 24 features are used to train the decision tree.

2.3 Training and Classification

We trained a decision tree using data collected from twelve different subjects including one female for 120 min of speaking and 120 min of silent data. We covered group conversation, public speaking, listening to class lecture, meeting, working at the work place for data collection scenarios. The training algorithm retains 80th percentile and standard deviation of inhalation, mean of exhalation, mean and median of IE ratio, mean and 80th percentile of B-duration, standard deviation of stretch as useful features. We observed classification accuracy of 93.08%, with Kappa = 0.8990, using decision tree (with pruning) with 10-fold cross validation. Using boosting (AdaBoostM1) the accuracy improves to 95.3571% with Kappa = 0.9303.

3. DEMONSTRATION

One of the the demonstrator will wear a RIP sensor which is connected with the band. In real-time, the phone will display the current state of the wearer whether (s)he is speaking or not. It will also show the respiration signal and features computed.

4. REFERENCES

[1] Android SDK. http://developer.android.com/sdk/index.html, 2010.

[2] AutoSense Project. http://sites.google.com/site/autosenseproject/, 2010.

[3] J. Ho and S. S. Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *ACM CHI*, 2005.

[4] G. Iachello and K. Truong. Prototyping and sampling experience to evaluate ubiquitous computing privacy in the real world. In *ACM CHI*, pages 1009–1018, 2006.

[5] P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring privacy concerns about personal sensing. *Pervasive Computing*, pages 176–183, 2009.

[6] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *ACM MobiSys*, pages 165–178, 2009.

[7] D. McFarland. Respiratory markers of conversational interaction. *Journal of Speech, Language, and Hearing Research*, 44(1):128–143, 2001.

[8] K. Plarre and A. Raij. Continuous Inference of Psychological Stress From Sensory Measurements Collected in the Natural Environment. In *IPSN*. ACM, 2011.

[9] Y. Wang and J. Lin. A framework of energy efficient mobile sensing for automatic user state recognition. In *ACM MobiSys*, pages 179–192, 2009.