How's my Mood and Stress? An Efficient Speech Analysis Library for Unobtrusive Monitoring on Mobile Phones

Keng-hao Chang, Drew Fisher, John Canny, and Björn Hartmann
Computer Science Division, University of California at Berkeley
Berkeley, CA, USA
{kenghao, dfisher, jfc, bjoern}@cs.berkeley.edu

ABSTRACT

The human voice encodes a wealth of information about emotion, mood, stress, and mental state. With mobile phones (one of the mostly used modules in body area networks) this information is potentially available to a host of applications and can enable richer, more appropriate, and more satisfying human-computer interaction. In this paper we describe the AMMON (Affective and Mental health MONitor) library, a low footprint C library designed for widely available phones as an enabler of these applications. The library incorporates both core features for emotion recognition (from the Interspeech 2009 Emotion recognition challenge), and the most important features for mental health analysis (glottal timing features). To comfortably run the library on feature phones (the most widely-used class of phones today), we implemented the routines in fixed-point arithmetic, and minimized computational and memory footprint. On identical test data, emotion and stress classification accuracy was indistinguishable from a state-of-the-art reference system running on a PC, achieving 75% accuracy on two-class emotion classification tasks and 84% accuracy on binary classification of stressed and neutral situations. The library uses 30% of real-time on a 1GHz processor during emotion recognition and 70% during stress and mental health analysis.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]: Real-time and embedded systems

General Terms

Algorithms, Design, Experimentation, Human Factors, Measurement, Performance

Keywords

Health Care, Mental Health, Monitor, Mobile Phones, Voice Analysis, Toolkit

1. INTRODUCTION

BodyNets 2011, Beijing, China, 7-8 November 2011, ISBN (978-1-936968-29-9).

Emotion, mood, stress, and mental health are key determinants of quality of life. *Affect* is a term used to cover mood and emotion. Mental health, especially depression, has close ties with emotion and e.g. is often first manifest as persistent negative mood. *Stress*, on the other hand, can be described as a "state of bodily or mental tension that tend to alter an existent equilibrium" [2]. Daily stress can have a considerable impact on individual mood and health.

Affective computing has a variety of applications: computers may adapt based on affect to improve learning, work performance, and communication [15]. Healthcare technologies can be made more intelligent to help people regulate emotions, manage stress, and avoid mental illness [18]. But capture of affect can be quite challenging, e.g. GSR sensors must be worn in the periphery of the body and primarily capture arousal, heart rate variability primarily captures stress and confounds with physical activity, and facial and body gesture conveys rich emotion but requires a camera pointing at the subject and real-time image analysis. On the other hand, voice is easily captured and has proved to be a surprisingly accurate tool for mental health evaluation, showing 90% classification accuracy for depression from a few minutes of voice data [17]. Voice analysis for emotion recognition [22] and stress detection [7] is somewhat less accurate but should be usable for everyday affect/mental health estimation (2-way emotion classification accuracies 70-80% and 4-way classification accuracy 61% for stress) .

Were one to design an ideal device for affect/mental health monitoring by voice, it would probably look a lot like a cell phone. A small, handheld device that is regularly used for voice-based tasks (i.e. calling others). A cell phone is arguably one of the mostly used modules in body area networks. What is lacking for developers are the speech features needed for applications or better still, binary or real values that denote emotion, stress or depression strengths - i.e. emotion classifier outputs.

While smartphones are gaining market share daily, "feature phones" are still the dominant devices in the hands of users, and will be for some time to come[1]. So to be feasible on feature phones and to be practical on smartphones, voice analysis must have a small computational footprint in both CPU time and memory. This is a primary goal in design of the AMMON library. The other goal is to ensure that analysis on the mobile library is as accurate as on a PC.

We have developed the AMMON library (Affective and Mental-health MONitor) to meet these goals. The library computes a rich set of prosodic and spectral features which support emotion recognition with state-of-the-art accuracy of around 70% based on the Interspeech 2009 emotion recognition reference dataset and feature

[1]Globally it seems unlikely that smartphones will ever dominate the market in developing countries

BODYNETS 2011, November 07-08, Beijing, People's Republic of China

DOI 10.4108/bodynets.2011.247079

set [22]. AMMON also includes features to describe glottal vibrational cycles, a promising feature for monitoring mental health. Moore et al. [17] showed that linear classifiers using a combination of these features can distinguish depressed and healthy subjects with 90% accuracy. This implies that the glottal activities in speech production can be greatly affected by mental illness, a good indicator of physical change induced by mental states. We hypothesize that the glottal features can improve stress detection as well. In analogy, mental stress often manifests physical response in the autonomic nervous system (cf. heart rate) [4]. So glottal features, indicating physical change in glottal muscles, may also respond to the autonomic nervous system. Our experiments showed that the glottal features indeed improved the classification accuracy. AMMON was written in C and we developed it based on an existing mobile front-end (ETSI advanced extended front-end [3]). AMMON will be available as open source, so researchers in the community can use it for various applications.

Most feature phones today lack floating-point hardware. They are almost all based on ARM9 processors. ARM's future roadmap for low-to-mid-range devices is based on the Cortex A5 family. Feature phones have clock speeds in the 150 to 400 MHz range. The toolkit we describe is intended to run on these feature phones. So far we have demonstrated 30% of real-time performance on 1GHz ARM devices using ARM9 instructions only, which should be close to real-time on 300MHz ARM devices[2].

The rest of this paper is structured as follows: Section 2 motivates the work by providing several sample applications. Section 3 describes the related work. Section 4 presents the speech analysis library, including the voice feature set and the improvement of algorithms for more efficient computation. It also includes the benchmarked performance running the library on mobile phones. Section 5 demonstrates the effectiveness of the features by applying them on an emotional speech dataset and a dataset of stress. The result matches the state-of-the-art result. Section 6 concludes the paper and discusses future work.

2. MOTIVATING APPLICATIONS

In this section we describe several applications that AMMON should be able to support, motivating the design requirement of the toolkit (Figure 1).

- Improving emotional intelligence. This application monitors a user's emotion continuously in order to improve the user's ability to identify, assess, and control their emotions. Even if users are good at assessing emotions over the short term, this application would allow visualization of frequency and intensity of emotions over the long term to expose trends in mood. By integrating contextual information like the user's calendar and location, the application can correlate emotions with possible triggers and allow the user to better manage those effects.

- Managing social relationships. This application would measure emotions and detect positive affect or conflicts during phone conversations. While users are generally aware of their emotions during a conversation, they are also cognitively loaded with the subject matter of the conversation. They may also fall without realizing into counterproductive roles (e.g. mutual victim roles in close relationships) which induce a variety of negative emotions or stress (frustration, defensiveness, anger) that are incorrectly attributed to the partner in the conversation. Emotion monitoring can help users better understand what they were actually feeling and expressing during a conversation with another.

- Computer-assisted psychotherapy. Almost all psychotherapies attempt to track patient's mental state in between therapy sessions. This includes mood, stress, triggers to emotion (the first bullet above), and direct cues to mental health. In conventional therapy this is limited to patient self-reports, which are often irregular and subject to a variety of biases. Monitoring of phone conversations should provide a more fine-grained and diverse sample. Beyond that, we are co-developing an "audio-diary" with researchers in San Francisco General hospital's psychiatry department. The diary elicits verbal records of patients mood during the day. It is less effortful than written records and provides a verbal transcript which complements the voice cues in the speech of the recording itself.

[2]The toolkit is not yet fully optimized, and e.g. does not yet use ARM intrinsics, so this figure should decrease.

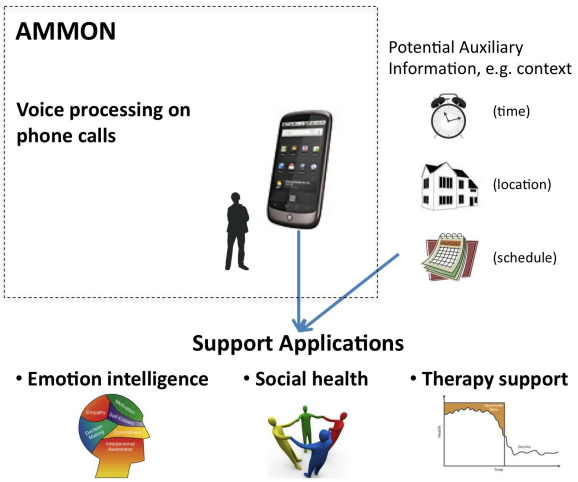

Figure 1: Illustration of AMMON applications. AMMON processes phone calls locally. Combined with with auxiliary information accessible to phones (e.g. context), it provides affective information to support a variety of applications

2.1 Design Requirement

The applications above require affect evaluation during continuous speech. Given the limited memory available in feature phones (typically a few Megabytes), it is not practical to buffer full conversational turns, and so real-time or near-real-time performance is important.

- Computational footprint. To make evaluation as accurate as possible, feature analysis should be as close to real-time as possible. Real-time is a moving target since processor speeds vary, but maximum reach would be achieved with real-time on/below 200 MHz processors. Memory on phone processors is important too and often neglected. The power required to drive the memory bus may actually dominate CPU power on small processors. Effective cache use can significantly reduce power consumption e.g. the SHRIMP system for camera-phone motion tracking achieved low power consumption in spite of real-time image processing by concentrating its computation in blocks that fit in the CPU's cache

[24].

- Optimal accuracy. Memory and CPU footprint can be reduced by naive means, e.g. by reducing speech sampling rates. But there is a heavy cost in accuracy for those changes. Instead we concentrate on simplifications that do not compromise accuracy. Our goal is to deliver an embedded, fixed-point affect estimation toolkit whose performance is indistinguishable from PC-based reference implementations.

3. RELATED WORK

In this section, we describe related work and discuss our contribution.

3.1 Emotion Recognition

Automatic emotion recognition has a long history from speech [13], and from facial expression and biosignals (e.g. heart rate and skin conductance) in [11]. For speech, an extremely useful landmark was the Interspeech Emotion Challenge 2009 [22]. This challenge included a standard dataset of emotion-tagged speech, and a "baseline" implementation of feature analysis, known as openSMILE. Surprisingly, while some more sophisticated algorithms improved on the baseline system, the improvements were very small, and it is fair to say that the baseline implementations achieved state-of-the-art performance. Since the baseline code was publicly distributed, we were able to compare our own implementation against it. A second surprising result was that use of segmental features (phone-level features) did not improve on "suprasegmental" primitive features (MFCCs, pitch, dynamics, energy). This may change in the future, but for now it means that state-of-the-art emotion recognition is much simpler than phonetic analysis. Expressed in terms of speech recognition components, that means that fully-accurate emotion analysis requires only the front-end of a speech recognizer and not the (memory and compute-intensive) acoustic model or later stages.

As a quick reference, the state-of-the-art recognition accuracy is about 70% for five-way classification of emotions (happy, sad, fear, anger and neutral) in a standard database with *actors* expressing emotions [19]. On the other hand, for the Interspeech challenge, *naturalistic* transcripts were recorded and hand annotated. Accuracy was only 70% for two-way classification [22].

3.2 Detecting Stress and Mental Health

A growing body of scientific research points towards psychomotor disturbances as consistent indicators (also known as prodrome [12]) of the onset of depression [20, 23]. It was also reported that psychiatrists routinely monitor these prodromes in patients during the diagnostic period and as measures for assessing treatment progress. For example, depressed patients often express slow responses (longer response time to questions and pause time within sentences), monotonic phrases (less fundamental frequency variability), and poor articulation. In a remarkable study, Moore et al. [17] showed that feature analysis can separate a control group of healthy subjects from a group of depressed patients with 90% accuracy. It relied most strongly on *glottal* features which are not part of most low-level speech analysis systems. We included these features in AMMON to support mental health analysis.

For stress detection, we applied AMMON to a publicly available stress speech dataset (SUSAS) [8]. As shown in Table 4 and 5, including glottal features led to improvement in the accuracy of stress classification. As a quick reference, Fernandez [7] performed a study to classify 4 stress categories during driving situations. The accuracy was 61% for 4-way classification.

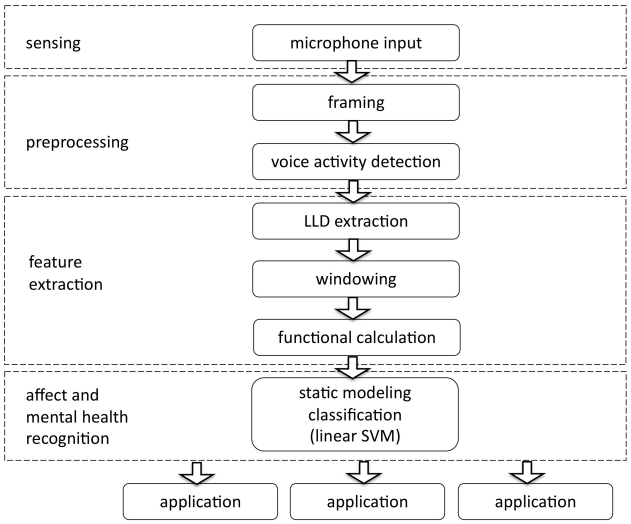

Figure 2: The AMMON Architecture

3.3 Voice Analysis Library on Mobile Phones

There has been a lot of activity lately on toolkits for mobile applications, including speech analysis and machine learning. SoundSense applied voice analysis to infer activities happening around a user, including driving, listening to music, and speaking [16]. SoundSense extracted a set of low-computation features and fed them to the J48 decision tree algorithm running locally on the phones. The features included zero-crossing rates, low energy frame rates, and other spectral features. By comparison, AMMON extracts affective features, including pitch and information about glottal vibrational cycles. It supports linear classification in real-time since the Interspeech challenge showed there to be little advantage in use of other classifiers for emotion recognition.

EmotionSense is an emotion recognition library on mobile phones for psychological studies [19]. EmotionSense does not infer emotions locally on the phones, but it ships the computation to the cloud. This imposes significant penalties in terms of privacy, need for access to the network, centralized server costs etc.

4. SPEECH ANALYSIS LIBRARY

In this section, we provide an overview of the AMMON architecture. We describe each architectural component in turn, as those illustrated in Figure 2.

Preprocessing. Sound processing starts with segmenting the audio stream from the microphone into frames with fixed duration (25 ms) and fixed stepping duration (10 ms). Not all frames are considered for further processing. The module performs *voice activity detection* for the non-speech frame dropping.

Feature Extraction. The selection of features is critical for building a robust classifier. We built a feature set based on the features defined in Interspeech challenge. It includes feature vectors derived by projecting *low-level descriptors* (LLDs, in the form of signal waveforms) such as pitch and energy by descriptive statistical *functionals* such as lower order moments (mean, standard deviation etc). The feature vectors were effective, which is probably justified by the supra-segmental nature of LDDs occurring with respect to the emotional content in speech [21].

Table 1 lists the LLDs in the categories of prosody, voice quality and spectral domains: zero-crossing rate (ZCR), root-mean-square

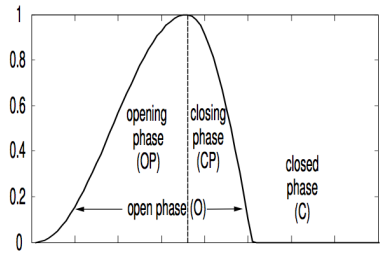

Figure 3: A Glottal Vibrational Cycle

(RMS) frame energy, pitch (F0), harmonics-to-noise ratio (HNR), mel-frequency cepstral coefficients (MFCC) 1-12. Moreover, for each of these LLDs, the delta coefficients are also computed.

In addition to the standardized set defined in the Interspeech challenge (16 LLDs), we include glottal timings in the LLDs, which had great success in measuring mental health [17]. As illustrated in Figure 3, a glottal (flow) vibrational cycle is characterized by the time that the glottis is open (*O*) (with air flowing between vocal folds), and the time the glottis is closed (*C*). In addition, an open phase can be further broken down into opening (*OP*) and closing (*CP*) phases. If there is a sudden change in airflow (i.e. shorter open and close phases), it produces more high frequency and the voice therefore sounds more *jagged*, other than *soft*. To capture it, AMMON calculates the above 4 durations of each cycle and 5 ratios of the closing to the opening phase (*rCPOP*), the open phase to the total cycle (*rOTC*), the closed phase to the total cycle (*rCTC*), the opening to the open phase (*rOPO*), and the closing to the open phase (*rCPO*). In summary, there were a total of 9 glottal timing-based LLDs included.

Then, AMMON segments the LLDs into windows, meaningful units for the modeling of feature vectors. A window can either be a turn or a fixed duration. Finally, it calculates 9 functionals from each window, including mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position and range. In the end, a feature vector contains $25 * 2 * 9 = 450$ attributes.

Affect and Mental Health Recognition AMMON uses linear Support Vector Machines (SVM) to recognize emotions based on the feature vectors (projecting LLDs by functionals). Linear SVM is currently a dominantly used mechanism for recognition of emotions. In addition, doing prediction with a linear SVM is rather efficient, which is suitable to run on the phones. Training models is more expensive, but this can be done off-line (not on the phones).

4.1 Implementation

We implemented AMMON in C, which can be deployed to both feature phones (e.g. Symbian) and smart phones (e.g. Android). In the paper we developed AMMON with Android NDK, where we can turn off the floating-point support in compile time to test the scenarios of feature phones. The Android platform has a dominant market share and is likely to lead the market in the near future. In addition, it supports implementation in both Java and C, which is convenient for re-using existing signal processing libraries written in C.

4.1.1 AMMON for Emotion Analysis

We developed AMMON by extending an ETSI (European Telecommunications Standards Institute) front-end feature extraction library [3]. The original purpose of the front-end was for local extraction of features on phones for remote speech recognition. The

Table 1: The AMMON feature set, computed by applying functionals on LLD waveforms.

LLDs	functionals
(Δ)ZCR (Δ)RMS energy (Δ)F0 (Δ)HNR (Δ)MFCC 1-12 (Δ)Glottal timings x 9 [a]	mean, standard deviation, kurtosis, skewness, minimum, maximum range, rel. position

[a]AMMON includes glottal timings for mental health anlysis, whereas the rest of LLDs are sufficient for emotion analysis.

ETSI front-end was useful for AMMON because (1) The ETSI front-end was already extracting some of the LLDs, such as energy, F0 and MFCC. We can re-use the code. (2) The front-end was equipped with noise-reduction routines, designed especially for the case of background noise while using mobile phones. It will make the features more reliable. (3) The library had routines for voice activity detection, which can be used for frame admission control. Non-speech frames will not be considered for further processing. (4) The ETSI library was implemented purely with fixed-point arithmetics, ensuring the library to run efficiently on feature phones without floating-point hardware.

After porting the front-end to Android, we implemented routines for the remaining LLDs (ZCR, HNR and glottal timings), using fixed-point arithmetic as well.

4.1.2 Extracting Glottal Timings

It is computationally more expensive to extract glottal timings than the other LLDs. So we implemented the routine with special care, including algorithmic improvement and code optimization. Following the algorithm proposed by Fernandez [6], we analyzed the bottleneck by profiling. The most dominant part is formant tracking, which requires for every sample, estimating LPC (linear predictive coding) polynomials and *solving roots* of each polynomial to determine formant frequencies. This part helps identify the closed-phases (C) of glottal vibrational cycles. When the glottis is closed, vocal tract is the only mechanism in effect in speech production. So formant frequencies should be stationary within short windows. In other words, we identify a closed phase if the formant frequency does not vary too much.

Solving roots of polynomials is expensive, which involves eigensolving the companion matrix of a polynomial. Even worse, the root solving is evoked frequenctly, in windows advancing in every sample. But we can leverage the property in a way to avoid constant eigensolving or "finding" roots from scratch. We can "track" roots instead. Because the polynomials are computed from adjacent windows that share a majority of speech samples, these LPC polynomials – and their roots – should not change a great deal between any two adjacent windows. Thus, we applied Newton-Raphson iteration to track roots of the current polynomial starting from the roots of the previous one. The Newton iteration is much cheaper. However, it does not guarantee to find all the roots. If it fails, we resort to the eigensolver, which always finds a correct answer but much more expensive. We applied several techniques to increase the probability of success in root tracking (e.g. subdivision between polynomials), but did it within the time budget that the Newton iteration gained over the eigensolver. We leave the details out here and will publish details in another report.

We implemented the Newton method ourselves, but for eigen-

solving, we applied CLAPACK [1]. However, the package was written in floating point. It is our future work to replace it with a fixed-point eigensolver, making AMMON truly applicable to feature phones (the remaining modules were implemented in fixed-point).

In addition to solving roots of the polynomials, the estimation of polynomials is also required to run in every sample. It involves using autocorrelation to construct *Toeplitz* matrices out of adjacent windows that share a majority of samples. We implemented the autocorrelation method in a way that the Toeplitz matrix is revised incrementally with each sample shift. This reduces the running time from quadratic to linear time.

The other bottleneck is the Fast Fourier Transform, which is evoked in every sample to calculate the phase change and locate the maximum excitation (the boundary between opening (OP) and closing (CP) phases). We improved this part with a piece of ARM optimized assembly code.

4.1.3 *Implementing Functionals*

Making a reliable estimate arguably requires as much data processing as possible. This means that, we have to calculate functionals over a large window of LLDs. Given the limited memory available on feature phones, it is not practical to buffer full conversational turns. AMMON should calculate the functionals over time without having to save the value at every sample. Therefore, we implemented an online, buffer-free algorithm to calculate the functionals (pseudo code can be found in [14]).

Given a new sample of an LLD, only the mean and the first to forth moments are updated, implying constant space per LLD. Then, it can calculate an up-to-date functional with the moments. For example, we can calculate variance with the second moment and obtain kurtosis with the forth and the second moments. In terms of computation, each update and computation of functionals takes only constant time.

4.2 Performance Evaluation

We evaluated the implementation in terms of its computational efficiency. And we break down the evaluation based on emotion recognition and mental health analysis.

4.2.1 *Emotion Analysis: Compare with openSMILE*

First, we compared AMMON with the open source toolkit used in the Interspeech challenge, named as openSMILE. For emotion recognition, we excluded the computation of glottal timings. Since AMMON has voice activity detection and noise suppression modules whereas openSMILE does not, we also intentionally turned them off for fair comparison.

As a benchmark, we made use of an emotional speech database (details in Section 5.1). There were 298 clips in the dataset, each with 10-60 seconds long. The benchmark was run on a Google Nexus One phone (1GHz Snapdragon CPU with floating-point hardware), where the floating-point was turned off to simulate the case of feature phones.

Table 2 shows that when the floating-point support was turned on (through compiler flags), AMMON ran comparably with openSMILE. OpenSMILE ran only slightly faster (17% of real time (xRT)) than AMMON (18% xRT), which supposedly was spending extra effort in fixed-point arithmetic. However, when the floating-point support was turned off, the fixed-point implementation paid off. OpenSMILE ran much slower (53% xRT), whereas AMMON stays the same (18% xRT). This implies that AMMON is more efficient than openSMILE on feature phones.

Finally, we turned on the modules of voice activity detection and noise suppression. AMMON ran in a total of 29% of real time. That said, AMMON can comfortably run emotion analysis in real time on 300MHz feature phones.

Table 2: Computational efficiency of AMMON. The running time are displayed in the percentage of real time (xRT) on a 1GHz phone.

toolkit	floating point ON	OFF
openSMILE	0.17 xRT	0.53 xRT
AMMON w/o Glottal Timings, VAD, and Noise Supr.	0.18 xRT	0.18 xRT

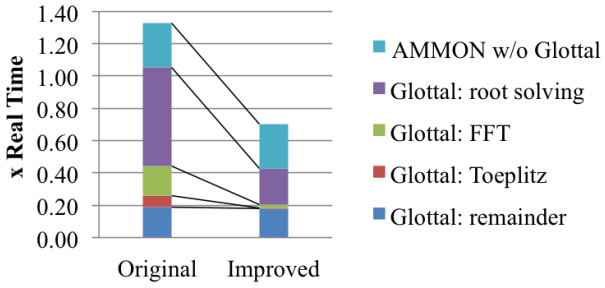

Figure 4: The breakdown of AMMON running time. The improvement of glottal extraction makes AMMON run 70% of real time on a 1GHz smartphone.

4.2.2 *Mental Health Analysis: Extract Additional Glottal Timings*

The modification described in Section 4.1.2 significantly improved the performance of glottal extraction, as illustrated in Figure 4. For root solving, we managed to reduce its running time by 68% (reduced to 1/3). Using assembly code for FFT reduced its running time by 85%. Incremental revision of the Toeplitz also reduced its running time in two orders of magnitude.

As a whole, the new glottal extraction algorithm ran from 105% of real time to 41% of real time, a 61% decrease. This adds up the AMMON computation time for mental health analysis to 70% of real time (was 133% of real time). That said, doing mental health analysis on phones is more expensive. AMMON can run mental health analysis on smart phones in real time, but about 2 times slower than real time over the feature phones. Nonetheless, in the next section we will show that the glottal features were indeed valuable, although it is computationally expensive. It significantly increased accuracy for stress detection. The value of glottal features for mental health analysis was already highlighted in [17].

5. FEATURE EVALUATION

In this section, we demonstrate the effectiveness of AMMON in the recogniton of emotions and mental stress. We will show that using the feature set extracted in AMMON, it recognizes emotions in state-of-the-art accuracy. We will also demonstrate that glottal features are helpful in seperating speech in stress from neutral conditions.

5.1 Emotion Recognition: The Belfast Emotional Dataset

Table 3: Comparison in the recognition of positive vs. negative emotional clips. We also list the F-measures for both clasess (class size: 112/133).

Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.778/0.727	0.753	75.51%
AMMON w/o Glottal timings	0.776/0.73	0.752	75.51%

We compared AMMON with the PC referencing system, i.e. openSMILE. As a benchmark, we could have chosen the FAU Aibo dataset used in the Interspeech challenge, where the recognition accuracy is available as a baseline for comparison. Nonetheless, given the goal of recognizing emotions in everyday conversations, the Aibo dataset is not entirely suitable. The Aibo dataset is in German, not in English. It is known that emotional expressions vary across languages and cultures, so a model trained in German may not be applicable to conversations in English. In addition, emotions happened in the database were mostly non-prototypical and subtle (e.g. empathy), making it insufficient to support most of the applications that require information of prototypical emotions (e.g. sad, happy etc).

Therefore, we chose the Belfast Naturalistic Database [5]. The dataset is in English, covering a wide range of emotional states that occur in everyday interactions, as well as prototypical examples of emotion such as full-blown anger. The Belfast database consists of 298 audiovisual clips from 125 speakers (31 males and 94 females). These clips were collected from a variety of television programs and studio-recorded conversations. Clips range from 10 to 60 seconds in length. The Belfast database were labeled by multiple raters. Each clip was labeled by 3 most visible emotions and the intensity (weak, medium and strong). We aggregated the labels in terms of voting and strength [10].

We performed a 2-way classification task to separate clips with positive emotions from those with negative emotions. The task is potentially useful for most applications, where the information of whether users are in positive or negative mood is of interest. A clip is considered positive if none of the label has negative valence, and vice versa. For the clips labeled with both positive and negative valence, we excluded them.

We applied AMMON to extract a feature vector from each clip. Note glottal timings were not extracted here. Then, we fed the feature vectors to SVM, a widely used method in emotion recognition (regularized linear SVM, features scaled, 5-fold cross validation). We applied the same procedure to openSMILE: extracting feature vectors and performing classification. There were a total of 112 positive valence clips (class 1) and 133 negative valence clips (class 2). Table 3 shows that AMMON had a comparable result to openSMILE, achieving 75% of accuracy and 0.75 ROC area. The accuracy resembles the result of the Interspeech challenge, around 70% in classifying 2 emotions in a naturalistic database. The experiment implies that AMMON can support emotion analysis in the same level of accuracy as the PC reference system, i.e. openSMILE.

5.2 Stress Detection: The SUSAS Stress Dataset

We evaluated stress detection with a dataset named Speech Under Simulated and Actual Stress (SUSAS) [8], developed by John Hanson. It is the most common dataset found in the literature for stress detection tasks [9]. For our experiment, we made use of the recordings under actual stress, where each subject was asked to speak (and repeat) 35 distinct English words while riding one of two roller coaster rides. High stress and neutral speech utterances were marked depending on the position of a riding course. There are a total of 7 subjects (3 females and 4 males) involved, producing a total of 1900 utterances. Each utterance was segmented as a word.

Table 4: Comparison in the recognition of stressed vs. neutral utterances, where AMMON includes additional glottal features (class size: 1200/701).

Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.867/0.770	0.763	83.18%
AMMON	0.877/0.788	0.832	84.43%

5.2.1 Recognizing Stressed vs. Neutral Speech

This was a 2-way classification problem, where utterances with high stress were put to class 1 and the neutral uttereances were assigned to class 2. We applied both AMMON and openSMILE to extract feature vectors, but this time, we included glottal features in AMMON. The feature vectors were fed into SVM (regularized linear SVM, features scaled, 10-fold cross validation). Table 4 shows that AMMON outperformed openSMILE with 1%, reaching 84% of accuracy (baseline is 63% because of the imbalanced data size 1200/701). Also, the ROC area significantly increased from 0.763 to 0.832. This demonstrates that, by adding glottal features AMMON can perform better in stress detection.

5.2.2 Recognizing Stress Increase vs. Stress Decrease in Speech

We hypothesized that stress detection can be further improved by user normalization. Because of user difference, the feature vectors in the previous task may be biased in different offsets in the feature space and ruin the classification. Nonetheless, if we look at the distance from a feature vector in neutral condition to another vector in stressed condition of the same user, we can focus on the within-user stress change (i.e. stress increase) and ignore the user difference.

Because of the nature of SUSAS, each user speaks the same set of words in both stress and neutral conditions. Therefore, we calculated the distance vector (by subtraction) between each pair of stress/neutral utterances of the same word by the same user. We also randomzed the order of subtraction so some distance vectors represent the increase of stress (a stress vector minus a neutral vector) whereas other distance vectors represent the decrease of stress.

The task became a two-way classification, where distance vectors with stress increase were put in class 1 and the distance vectors with stress decrease were put to class -1. Note we included distance vectors by all users in the same pool, so this is a user-independent classifier. We applied both AMMON and openSMILE to extract feature vectors. Similarly, we also included glottal features in AMMON. The feature vectors were fed into SVM (regularized linear SVM, features scaled, 10-fold cross validation). Table 5 shows that AMMON outperformed openSMILE with 1.3%, reaching 93.6% of accuracy (baseline is 50% because of the dataset is symmetric and balanced). The 1.3% increase is significant at the 92% accuracy level. Also, the ROC area increased from 0.923 to 0.936. This again demonstrates that, by adding glottal features AMMON can perform better in stress detection. Glottal features can be another way of refelcting physical response to stress in the human voice.

Readers may be questioning that this is not the real stress vs.

Table 5: Comparison in the recognition of stress increase vs. stress decrease, where AMMON include glottal features (class size: 337/336)

Feature Set	F-Measures	ROC Area	Accuracy
openSMILE	0.923/0.923	0.923	92.27%
AMMON	0.936/0.936	0.936	93.60%

neutral classification. Nonetheless, we argue that this result is probably more useful in real world applications. We can calculate the difference from the current feature vector to the next, and judge whether a user has increased or decreased the stress level. In addition, the result is very promising (93% accuracy for a balanced dataset), and can be used to create a temporal model of stress detection.

6. CONCLUSION

The pervasiveness of mobile phones opens up an opportunity for improving our psychological well-being, and it scales from individuals to the mass public. An emotion monitor can raise individual awareness and contribute behavior change, and a stress detecter can help individuals manage their stress. Beyond this, a mental health tracker can detect early stage problems, measure health trend of the public, and promote public health. Therefore, we propose AMMON, an affective and mental health monitor. AMMON was designed to work on feature phones, so that most people can have access to this service. We were able to prove that the features extracted by AMMON were as effective as those by reference systems on PC. AMMON can recognize emotions in state-of-the-art accuracy and detect stress with improved accuracy by the additional glottal features. The glottal features will also benefit mental health analysis, e.g. depression. We will open source this library, but before that, we will optimize it by using ARM intrinsics, making it run faster and put less burden on phone processors. In addition, we will replace the floating-point eigensolver library with a fixed-point version. We vision that this toolkit will be able to support a host of research projects that require efficient affect recognition capability.

7. REFERENCES

[1] CLAPACK (f2c'ed version of LAPACK). http://www.netlib.org/clapack/, retreived in May 2011.

[2] Stress. in merriam-webster's medical dictionary. merriam-webster, inc., 2007. Available: http://dictionary.reference.com/browse/stress. Accessed: September 07, 2011.

[3] ES 202 212 extended advanced front-end feature extraction algorithm v1.1.4. Technical report, ETSI, 2005. Source code retrievable through secure login on http://www.etsi.org/WebSite/Technologies/DistributedSpeechRecognition.aspx, May 2011.

[4] J. Choi and R. Gutierrez-Osuna. Using heart rate monitors to detect mental stress. In *IEEE Body Sensor Networks*, 2009.

[5] E. Douglas-Cowie, R. Cowie, and M. Schroeder. The description of naturally occurring emotional speech. In *15th ICPhS*, 2003.

[6] R. Fernandez. *A Computational Model for the Automatic Recognition of Affect in Speech*. PhD thesis, MIT, 2004.

[7] R. Fernandez and R. W. Picard. Modeling drivers' speech under stress. *Speech Communication - Special issue on speech and emotion*, 40(1-2), 2003.

[8] J. H. L. Hansen. Susas. Linguistic Data Consortium, Philadelphia, 1999.

[9] J. H. L. Hansen and S. Patil. *Speech under stress: Analysis, modeling and recognition*, volume 4343. SpringerLink, 2007.

[10] K. hao Chang and J. Canny. Lessons learned in modeling vocal expression of affect from a naturalistic emotional database. Technical report, UC Berkeley, 2011.

[11] J. Healey, J. Seger, and R. Picard. Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations. In *the Rocky Mountian Bio-Engineering Symposium*, 1999.

[12] A. Jackson, J. Cavanagh, and J. Scott. A systematic review of manic and depressive prodromes. *J Affect Disord.*, 74(3):209–217, 2003.

[13] P. Juslin and K. Scherer. *Vocal expression of affect*, chapter 3, pages 65–135. Oxford University Press, 2005.

[14] D. E. Knuth. *The Art of Computer Programming*, page 232. Boston: Addison-Wesley, 1998.

[15] B. Kort, R. Reilly, and R. W. Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy building a learning companion. In *ICALT-2001 (International Conference on Advanced Learning Technologies)*, 2001.

[16] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. SoundSense: Scalable sound sensing for people-centric sensing applications on mobile phones. In *Proc. of 7th ACM Conference on Mobile Systems, Applications, and Services (MobiSys '09)*, 2009.

[17] E. Moore, M. Clements, J. Peifer, and L. Weisser. Comparing objective feature statistics of speech for classifying clinical depression. In *IEMBS*, 2004.

[18] M. E. Morris, Q. Kathawala, T. K. Leen, E. E. Gorenstein, F. Guilak, M. Labhard, and W. Deleeuw. Mobile therapy: Case study evaluations of a cell phone application for emotional self-awareness. In *J Med Internet Res*, 2010.

[19] K. K. Rachuri, P. J. Rentfrow, M. Musolesi, C. Longworth, C. Mascolo, and A. Aucinas. EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. In *Ubicomp*, 2010.

[20] D. Schrijvers, W. Hulstijn, and B. G. Sabbe. Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool. *J Affect Disord*, 109:1–20, 2008.

[21] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classiPcation of emotional user states: Low level descriptors and functionals. In *Interspeech*, 2007.

[22] B. Schuller, S. Steidl, A. Batliner, and F. Jurcicek. The interspeech 2009 emotion challenge: Results and lessons learnt. http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2009-10/interspeech-emotion-challenge/, retrieved in May 2011, 2009.

[23] C. Sobin and H. A. Sackeim. Psychomotor symptoms of depression. *Am J Psychiatry*, 154:4–17, 1997.

[24] J. Wang, S. Zhai, and J. Canny. SHRIMP: solving collision and out of vocabulary problems in mobile predictive input with motion gesture. In *CHI*, 2010 ACM Conference on Human Factors in Computing Systems (CHI'10).