AMMON: A Speech Analysis Library for Analyzing Affect, Stress, and Mental Health on Mobile Phones

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

Abstract

The human voice encodes a wealth of information about emotion, mood and mental state. With mobile phones this information is potentially available to a host of applications. In this paper we describe the AMMON (Affective and Mental-health MONitor) library, a low footprint C library designed for widely available phones. 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 most of the routines in fixed-point arithmetic, and minimized computational and memory footprint. While there are still floating-point routines to be revised in fixed-point, on identical test data, emotion and mental stress classification accuracy was indistinguishable from a state-of-the-art reference system running on a PC.

1 Introduction

Emotion, mood 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. Affective computing has a variety of applications: computers may adapt based on affect to improve learning, work performance etc. Healthcare technologies can be made more intelligent to help people regulate emotions, manage stress, and avoid mental illness. But capture of affect can be quite challenging, e.g. GSR and heart rate sensors must be worn in the periphery of the body. On the other hand, voice is easily captured and has proved to be a surprisingly accurate tool for mental health evaluation, e.g. showing 90% classification accuracy for depression from a few minutes of voice data [10]. Voice analysis for emotion recognition [12] is somewhat less accurate (accuracies 70-80%) but should be usable for everyday affect/mental health estimation.

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). What is lacking for developers are the speech features needed for applications or better still, binary or real values that denote emotion 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 set [12]. AMMON also includes features to describe glottal vibrational cycles, a promising feature for monitoring depression. Moore et al. [10] 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) [3]. 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 [2]). 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. 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 and 45-65% of real-time on 600 MHz ARM devices, which should be close to real-time on 300-400MHz ARM devices[2].

The rest of this paper is structured as follows: Section 2 describes the related work. Section 3 presents the speech

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

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

analysis library, including the voice feature set and the effort to improve efficiency. It includes the benchmarked performance running the library on mobile phones. Section 4 demonstrates the effectiveness of the features by applying them on an emotional speech dataset and a dataset of mental stress. The result matches the state-of-the-art result. Section 5 concludes the paper and discusses future work.

2 Related Work

Automatic emotion recognition has a long history wth speech processing [7]. An extremely useful landmark was the Interspeech Emotion Challenge 2009 [12]. This challenge included a "baseline" implementation of feature analysis, known as openSMILE. Since the baseline code was publicly distributed, we were able to compare our own implementation against it.

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 [9]. 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 [11]. 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.

3 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 1.

Preprocessing. Sound processing starts with segmenting the audio stream from the microphone into frames with fixed duration (200 ms) and fixed stepping duration (80 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 static 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).

Table 1 lists the LLDs in the categories of prosody, voice quality and spectral domains: zero-crossing rate (ZCR), root-mean-square (RMS) frame energy, pitch (F0), harmonics-to-noise ratio (HNR), mel-frequency cepstral coefficients (MFCC) 1-12. Moreover, to each of these LLDs, the delta coefficients are additionally computed.

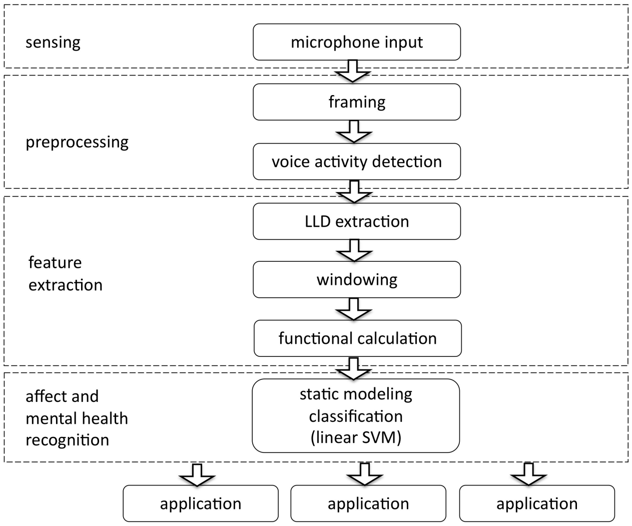

Figure 1. The AMMON Architecture

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 [10]. 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). 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).

3.1 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.

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.

3.1.1 AMMON for Emotion Analysis

We developed AMMON by extending an ETSI (European Tele-communications Standards Institute) front-end feature extraction library [2]. The original purpose of the front-end was for local extraction of features on phones for remote speech recognition. Nonetheless, the 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 in particular.

3.1.2 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 algorithimic improvement and code optimization. Following the algorithm proposed by Fernandez [5], 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.

Solving roots of polynomials is expensive, which involves eigensolving the companion matrix of a polynomial. Even worse, the root solving is evoked frequently, 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 but 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 eigensolving, 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 optimized the part with a piece of ARM optimized assembly code.

3.1.3 Implementing Functionals

Making a reliable estimate arguably requires as much data processing as possible. 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 [8]).

3.2 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.

3.2.1 Emotion Analysis: Compare with openSMILE

First, we compared AMMON with the open source toolkit openSMILE used in the Interspeech challenge. 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 4.1). 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

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

(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. We also benchmarked the performance on two slower phones with 600 MHz CPU (Motorola Droid with TI OMAP 3430 CPU and HTC Aria with Qualcomm MSM7227 CPU). AMMON ran in a total of 45% of real time on Motorola Droid and 64 % of real time on HTC Aria. That said, AMMON should run emotion analysis in real time on 300-400 MHz feature phones [3].

3.2.2 Extract Additional Glottal Timings

Since the root solving module was not revised to fixed-point yet, we turned on the floating point support for the extraction glottal features. The modification in Section 3.1.2 significantly improved the performance of glottal extraction, as illustrated in Figure 2. 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 are 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 recognition accuracy for mental health analysis.

4 Feature Evaluation

In this section, we demonstrate the effectiveness of AMMON in the recognition of emotions and the monitor of mental stress.

4.1 Emotion Recognition: The Belfast Dataset

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. 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). Instead, we chose the Belfast Naturalistic Database [4]. 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.

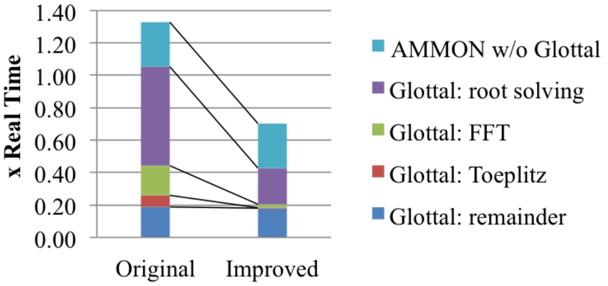

Figure 2. The breakdown of AMMON running time. The improvement of glottal extraction makes AMMON run 70% of real time on a 1GHz smartphone.

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. We excluded the clips labeled with both positive and negative valence.

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.

4.2 Stress Detection: The SUSAS Dataset

We evaluated stress detection with a dataset named Speech Under Simulated and Actual Stress (SUSAS). It is the most common dataset found in the literature for stress detection tasks [6]. 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 rid-

[3] Extrapolation based on by CPU frequency scaling may not hold due to factors such as slower and smaller memory systems, so benchmark will be made on more phones as a future work.

Table 3. Comparison in the recognition of positive v.s. negative emotional clips. We also list the F-measures for both classes (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%

Table 4. 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%

ing 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. Each utterance was segmented by a word, averaging about 1 second.

Previous work showed that user difference is significant in the vocal expression of mental stress. The user difference may bias feature vectors in the feature space differently and ruin the classification. Therefore, we focused on the "change" of feature vectors. We wanted to understand whether the mental stress alters and shifts the features in the same manner (i.e. in terms of vector direction and magnitude) across all users in the feature space. So we calculated the distance vector (by subtraction) between each pair of stress/neutral utterances of the same word by the same user. We also randomized the order of subtraction so some distance vectors represent stress increase (a stress vector minus a neutral vector; class 1) whereas other distance vectors represent stress decrease (class -1), becoming a two-way classification.

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 1.3%, 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 demonstrates two things. First, the features extracted by AMMON can identify the stress change on features very well (93% accuracy for a balanced dataset). Second, the glottal features, which proved to be helpful in the detection of depression, can improve the classification of mental stress. It reflects the physical response to stress in the human voice.

5 Conclusion

In this paper 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 analyze mental stress with improved accurarcy by the additional glottal features. 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 are investigating ways to replace the floating-point eigensolver library with a fixed-point version. However, re-inventing a fixed-piont eigensolving library is not trivial. We are also considering the Jenkins-Traub algorithm to replace the companion matrix/eigensolving method for root solving.

6 References

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

[2] 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, May 2011.

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

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

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

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

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

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

[9] 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.

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

[11] 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.

[12] B. Schuller, S. Steidl, A. Batliner, and F. Jurcicek. The interspeech 2009 emotion challenge: Results and lessons learnt. SLTC Newsletter, October 2009.