The SRI AVEC-2014 Evaluation System

Vikramjit Mitra
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
vikramjit.mitra@sri.com

Elizabeth Shriberg
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
elizabeth.shriberg@sri.com

Mitchell McLaren
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
mitchell.mclaren@sri.com

Andreas Kathol
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
andreas.kathol@sri.com

Colleen Richey
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
colleen.richey@sri.com

Dimitra Vergyri
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
dimitra.vergyri@sri.com

Martin Graciarena
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
martin.graciarena@sri.com

ABSTRACT
Though depression is a common mental health problem with significant impact on human society, it often goes undetected. We explore a diverse set of features based only on spoken audio to understand which features correlate with self-reported depression scores according to the Beck depression rating scale. These features, many of which are novel for this task, include (1) estimated articulatory trajectories during speech production, (2) acoustic characteristics, (3) acoustic-phonetic characteristics and (4) prosodic features. Features are modeled using a variety of approaches, including support vector regression, a Gaussian backend and decision trees. We report results on the AVEC-2014 depression dataset and find that individual systems range from 9.18 to 11.87 in root mean squared error (RMSE), and from 7.68 to 9.99 in mean absolute error (MAE). Initial fusion brings further improvement; fusion and feature selection work is still in progress.

Categories and Subject Descriptors
I.5.4 [Computing Methodologies]: Pattern Recognition– *signal processing*; *Waveform analysis*
J.4 [Computer Applications]: Social and Behavioral Sciences– *psychology*
G.3 [Mathematics of Computing]: Probability and Statistics– *correlation and regression analysis, robust regression, time series analysis*;

AVEC'14, November 07 2014, Orlando, FL, USA
Copyright 2014 ACM 978-1-4503-3119-7/14/11…$15.00
http://dx.doi.org/10.1145/2661806.2661818

General Terms
Algorithms, Measurement, Performance, Design, Reliability, Experimentation, Verification.

Keywords
Depression, robust signal analysis, acoustic features, articulatory features, prosody, support vector regression, decision trees, time series prediction.

1. INTRODUCTION
Depression affects the psychological state of a wide range of the human population and can be life-threatening for men, women [1] and even children [2]. Detection, evaluation, early treatment and therapy [3, 4] can help save a significant portion of patients suffering from depressive disorders and improve their quality of life. Accurate diagnosis of depressive symptoms, mainly assessed during interviews between the patient and the clinician, requires intensive training, experience and time. These clinical assessments provide the patient with an objective score based on the combined observation [21] of key symptoms typically observed with depression. Such interview-driven diagnosis is subjective in nature and both labor- and time-intensive. Automatic detection of depression can help physicians and medical practitioners detect depression earlier and facilitate quick evaluation and treatment. Audio-visual data can be used to build automated systems for depression detection, where knowledge of the key bio-signatures of depression in those data plays a dominant role in assuring high accuracy.

Studies [5, 6] have analyzed the speech patterns of depressed patients before and after treatment with antidepressant medication; other studies [7, 8] have reported that the speech of a depressed person undergoes a shift compared to non-depressed subjects. Detection of depression from speech has been explored by several researchers and research groups in the past. The Audio-Visual Emotion recognition Challenge (AVEC) [9], by providing audio-visual data for researchers, has created an ideal platform to

develop and evaluate systems for automated depression and affect detection.

Speech has been one of the prime modalities explored for depression detection. A wide array of features have been explored in the literature, starting with standard mel-cepstral features (commonly known as the MFCCs) [10, 11], prosodic features (such as pitch, energy, and speaking rate, etc.) [12, 13, 14], and traditional spectral-based features (such as formants, formant bandwidths, spectral energies, spectral tilt, etc.) [11, 12, 13, 14, 15, 22]. Recently, correlation structure features have been proposed in [16] along with delta mel-cepstral features and formant trajectories; using these features, significant improvement in depression level detection was observed in comparison to the standard baseline system. In a different study [17], MFCCs and their velocity and acceleration coefficients were used along with advanced machine learning techniques to perform depression level detection. Several studies [17, 18, 19, 20] have used both audio and video modalities. [18, 19] demonstrated that use of both the modalities improves the accuracy of an automated system compared to using each modality by itself. [17] demonstrated that the audio modality can give slightly better results than video; however, their combination performed slightly worse than both of them.

In this paper we present our system, which is designed for the depression detection sub-challenge (DSC) of the AVEC-2014 challenge [23] workshop using only audio data. The goal of this challenge is to predict, using audio-video data, an individual's self-reported depression score, specified according to the Beck depression rating scale [24]. We analyzed an array of acoustic features that capture key relevant signatures of depression from speech and compared their individual performance with respect to each other. We present a robust fusion of multiple systems, which exploits complementary information amongst the subsystems to produce a robust prediction of depression score from speech. The detailed analysis provided us with key insights regarding the depression signatures in each of the features we explored and helped us to produce a system that outperforms the baseline system [23].

The rest of the paper is organized as follows. In Section 2 we present the audio-visual AVEC 2014 challenge data. Section 3 presents the audio features explored in our work. Section 4 presents the machine learning algorithms we explored. Section 5 presents the results, followed by conclusions and future directions in Section 6.

2. AVEC DATABASE

The AVEC-2014 challenge dataset is a subset of the AVEC 2013 audio-visual depression corpus [9], which contains 150 videos of subjects performing a human-computer interaction task while being recorded by a webcam and a microphone. Each recording consists of only one person. The total number of subjects in the entire dataset is 84. Some subjects were recorded more than once: 18 subjects appear in three recordings, 31 in two, and the remaining 34 in only one recording. The duration of each recording ranged from 20 minutes to 50 minutes with an average duration of 25 minutes. The total duration of all clips is 240 hours. The average age of subjects was 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 63 years. The recordings took place in a number of quiet settings; however, we observed some ambient noise, reverberation and artifacts introduced by the background into the audio part of the recordings.

The recordings consisted of sustained vowel phonations which were spoken out loud while solving a task: counting numbers from one to ten; reading out loud; singing; telling a story from the subject's own past; and telling an imagined story. The recordings in the AVEC-2014 subset consist of only two tasks [23]: Northwind and Freeform, which were supplied as separate recordings, resulting in a total of 300 videos. The set of source videos is largely the same as that used for the AVEC-2013 challenge; however, five pairs of previously unseen recordings were used by the organizers to replace a small number of videos used in the 2013 challenge.

The audio data was collected using a headset microphone connected to the sound card of a laptop and sampled at various sampling rates. We re-sampled all the data to 16 kHz. The challenge data was split into three partitions of training, development and test sets with 50 Northwind-Freeform pairs in each set for a total of 300 task recordings. The train, dev and test sets had similar distributions in terms of age, gender, and depression levels for the partitions. There was no session overlap between partitions. The target depression scores for the training and development set were distributed to the challenge participants by the organizers. The test set scores were not provided. The estimated test set scores had to be sent to the challenge organizers who performed the scoring of the system performance in terms of mean absolute error (MAE) and root mean squared error (RMSE).

3. AUDIO FEATURES

We explored a wide array of acoustic features that capture speech articulation, acoustic-phonetic information, spectral representation, speech modulation, vocal effort, rhythmicity, speech prosody, vowel stress, speech intensity, etc. These features operate at multiple scales: some are low-level descriptors computed at a specific frame rate, while others are global descriptors, i.e., one feature for the whole waveform. We restrict ourselves to automatically extractable features that do not rely on words for two reasons: privacy and practicality. Word features also require speech recognition, which may or may not be available at high enough performance levels for a particular individual or context. The details of each of the features explored in our work are provided below.

Damped Oscillator Cepstral Coefficients (DOCC) [25] aim to model the dynamics of the hair cells within the human ear and try to capture the perceptually relevant information from audio. In the human auditory system, the hair cells detect the motion of incoming sound waves and excite the neurons of the auditory nerves. In DOCC processing (shown in Figure 1), the incoming speech signal is analyzed by a bank of gammatone filters (in this work, we used a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale), which splits the signal into band-limited subband signals. In turn, these subband signals are used as the forcing functions to an array of damped oscillators whose response is used as the acoustic feature.

As shown in Figure 1, each band-limited time signal from the gammatone filterbank is used to excite a forced underdamped oscillator and the response of the oscillator is given as:

$$x[n] = \frac{(2\zeta\Omega_0^2)F_e[n] + 2(1+\zeta\Omega_0)x[n-1] - x[n-2]}{(1+2\zeta\Omega_0+\Omega_0^2)} \quad (1)$$

$$\Omega_0 = \omega_0 T \text{ and}$$
$$T = 1/f_s$$

where F_e is the externally applied force on the oscillators, in this case the output of the gammatone filterbank. ω_0 is the undamped angular frequency of the oscillator; ζ is the damping ratio; and f_s is the sampling rate.

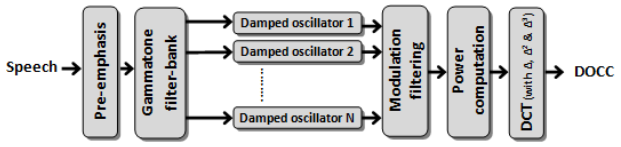

Figure. 1. Block diagram of the damped oscillator-based feature extraction.

DOCC is a channel- and noise-robust acoustic feature and, given that the AVEC-2014 audio data contain some degree of background distortion, it can be expected that robust features will be more effective than traditional MFCCs, which are more susceptible to noise and background distortions. Figure 2 shows the FFT spectrum of a signal corrupted with 3 dB high pass noise and the damped oscillator response of the same signal.

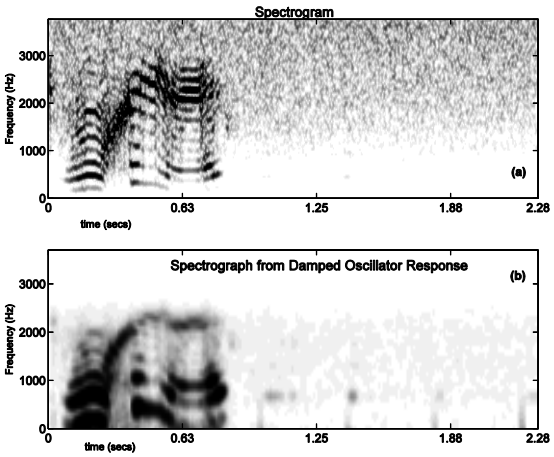

Figure. 2. (a) Spectrogram of signal corrupted with 3 dB noise (top panel) and (b) Spectral representation of the damped oscillator response (bottom panel).

More details about damped oscillator processing and the DOCC pipeline can be obtained in [25]. We analyzed the damped oscillator response by using a Hamming analysis window of 26 ms with a frame rate of 10 ms. The power signal from the damped oscillator response was computed, then root compressed using the 15th root followed by discrete cosine transformation (DCT). We retained the first 13 DCT coefficients and used them as the DOCC feature vector in our experiments.

Normalized Modulation Cepstral Coefficients (NMCC) [26] are perceptually-motivated noise-robust acoustic features which are based on the speech perception studies [27, 28], which state that amplitude modulation (AM) of subband speech signals plays a pivotal role in human speech perception and recognition. Figure 3 shows the block diagram of NMM feature generation.

The NMCCs are obtained by tracking the AM trajectories of subband speech signals in the time domain using a Hamming window of 26 ms with a frame rate of 10 ms. In this processing, the speech signal was analyzed using a time-domain gammatone filterbank with 34 channels equally spaced on the ERB scale. The subband signals from the gammatone filterbanks were then processed using the Discrete Energy Separation algorithm (DESA) [29], which produced instantaneous estimates of AM signals. The powers of the AM signals were then root compressed using the 15th root and their DCT coefficients were generated. From these, only the first 13 coefficients were selected for use in the NMCC feature vector in our experiments. NMCCs have been found to be more noise- and channel-robust than MFCCs for automatic speech recognition (ASR) [26], speaker identification [30] and language recognition [31].

Figure 3. Flow diagram of NMCC feature extraction from speech.

Modulation of Medium Duration Speech Amplitudes (MMeDuSA) [32, 33] features aim to track the subband AM signals of speech, but they use a medium duration analysis window and also track the overall summary modulation. The summary modulation plays an important role in tracking speech activity and in locating events such as vowel prominence/stress, etc. Figure 4 shows the block diagram of the MMeDuSA feature generation pipeline.

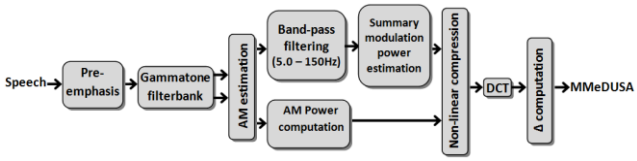

Figure 4. Flow diagram of MMeDuSA feature extraction from speech.

The MMeDuSA feature generation pipeline uses a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. It employs the nonlinear Teager energy operator [34] to crudely estimate the AM signal from the band-limited subband signals. MMeDuSA uses a medium duration Hamming analysis window of 52 ms with a 10 ms frame rate and computes the AM power over the analysis window. The powers were root compressed using the 15th root and their DCT coefficients were obtained. From these, the first 13 coefficients were retained. Additionally, the AM signals from the subband channels were bandpass filtered to retain the modulation information within the 5 to 200 Hz range, which was then summed across the frequency scale to produce a summary modulation signal. The power signal of the modulation summary was obtained, followed by 15th root compression. The result was transformed using DCT and the first three coefficients were retained and combined with the previous 13-dimensional features to produce 16-dimensional MMeDuSA features.

Gammatone Cepstral Coefficients (GCCs) use the gammatone filters, which are a linear approximation of the auditory filtering performed in the human ear. In GCC processing, speech is analyzed using a bank of 40 gammatone filters equally spaced on the ERB scale. The power of the bandlimited time signals within an analysis window of 26 ms was computed at a frame rate of 10 ms. Subband powers were then root compressed using the 15^{th} root and DCT was performed on the resultant. The first 13 DCT coefficients were retained as the GCC feature vector.

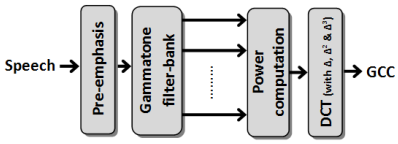

Figure 5. Flow diagram of GCC feature extraction from speech.

Articulatory Features (AFs), articulatory motions from spontaneous speech, have been demonstrated by previous studies [35, 36] to provide a sufficient degree of robustness for speech recognition tasks. Because depression affects a speaker's production system, these features can potentially capture the relevant signatures of depression from speech. In this work, we used a deep neural network (DNN) with 150, 200, 100, 80, 60, and 40 neurons [35], where the number of neurons in each layer was optimized empirically and the depth of the network was increased incrementally. In this DNN architecture the input observations were time-contextualized NMCC features (generated from the acoustic waveform with multiple frames concatenated across time) and the outputs were time-domain vocal tract constriction variables (also abbreviated as TVs) as shown in Figure 7; their details are provided in Table 1. Figure 6 shows the articulatory feature extraction pipeline.

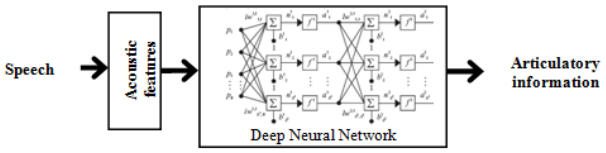

Figure 6. Flow diagram of articulatory feature extraction from speech.

Due to the lack of a natural speech dataset containing parallel data of acoustic waveforms and TVs, we used Haskins Laboratories' Task Dynamic model (known as TADA [37]) to generate a synthetic English isolated-word speech corpus along with the TVs. TADA was used to generate a synthetic word corpus of 111,929 words, where the words were borrowed from the CMU dictionary. TADA generated the corresponding TVs, which are eight vocal tract constriction variables corresponding to: Lip Aperture (LA); Lip Protrusion (LP); Tongue Tip Constriction Degree (TTCD); Tongue Tip Constriction Location (TTCL); Tongue Body Constriction Degree (TBCD); Tongue Body Constriction Location (TBCL); Velic Opening (VEL); and Glottal Opening (GLO). 80% of the synthetic data was used for training the DNN; 10% was used as the cross-validation set; and the remaining 10% was used to test the DNN.

Table 1. Constriction organ, vocal tract variables, their unit of measurement and dynamic range.

Constriction organ	Vocal tract variables	*Unit*	*Dynamic range*	
			Max	*Min*
Lip	Lip Aperture (LA)	mm	27.00	-4.00
	Lip Protrusion (LP)	mm	12.00	8.08
Tongue Tip	Tongue tip constriction degree (TTCD)	mm	31.07	-4.00
	Tongue tip constriction location (TTCL)	degree	80.00	0.00
Tongue Body	Tongue body constriction degree (TBCD)	mm	12.50	-2.00
	Tongue body constriction location (TBCL)	degree	180.00	87.00
Velum	Velum (VEL)	-	0.20	-0.20
Glottis	Glottis (GLO)	-	0.74	0.00

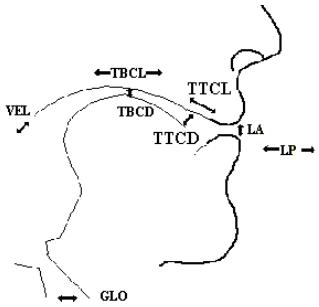

Figure 7. Eight tract variables from five distinct constriction locations.

Acoustic Phonetic (AP) features [43] represent acoustic-phonetic information (e.g., formant information, mean Hilbert envelope, periodic and aperiodic energy in subbands [44], etc.) and are analyzed at a 5 millisecond frame rate with a 10 ms analysis window. 13 APs were selected to represent information such as reflection coefficients, mean Hilbert envelope, periodic energy, aperiodic energy [44], nasal energy [45], etc. Data such as periodic energy, Hilbert envelope, etc. provide information regarding voice quality, energy contour, etc., which should help to capture the level of emotion in speech.

The **Kaldi Pitch** tracker [41] comes with the Kaldi pitch recognition toolkit [42] and provides a two-dimensional output consisting of pitch tracks and a normalized cross-correlation function that gives indication about voicing information. Depression usually results in speech with a lesser degree of excitement and pitch tracks would help to capture the degree of excitement in speech.

Energy contour features: we also computed features intended to capture longer range information and/or simple statistics over a recording. The energy contour (**encon)** feature [38] set aims to capture rhythmicity as well as overall speaking rate (without relying on phone recognition) by looking at the periodicity of

energy peaks within each segment. The motivation for this work is that depressed speech may have a lower overall rate and be more temporally monotonous. This feature models the contour of 10 millisecond windows of the first two coefficients (c_0 and c_1) from an MFCC front end; each cepstral stream is mean-normalized over the utterance, making it robust to absolute level differences over both entire sessions and within-session segments. A discrete cosine transform is then taken over a 200 ms sliding window with a 100 ms shift. Vector components comprise the first 5 and 2 bases from the DCT over each window of c_0 and c_1, respectively.

Spectral tilt features aim to capture vocal effort in a manner quasi-robust to extrinsic session variability, using methods developed in [38]. These features were extracted for voiced frames. Voicing was determined using a logistic regression classifier trained with number of zero crossings, log energy, number of peaks in the autocorrelation of the window signal, and standard deviation of the inter-peak distance, where the voicing threshold was set to 0.5. The five component spectral tilt features include H2-H1, F1-H1, and F2-H1 (where H1, H2 are the lower-order harmonics and F1,F2 are the first two formants), that reflect lower-order harmonics and formants given the microphone and room conditions. The last two features are measures of the spectral slope per frame, and the difference between the maximum of the log power spectrum and the maximum in the 2kHz-3kHz range.

DLE features were designed to capture very local vocal effort changes in a manner that does not require normalization for overall speaking level. Our hypothesis is that such features can help to capture the degree of prosodic accentuation, which we expect to be lower for depressed subjects. These features measure the difference in log energy locally at the transition frames from voiceless speech to voiced speech, and conversely from voiced to voiceless speech. DLE features are thus quite sparse, occurring only once per voiced region and with only a single feature dimension.

Intonation-related features include **Pitch** features: f0, f0pk and f0pk-stats), which capture frame-level pitch, pitch peak distributions, and various statistics on the location of pitch peaks relative to each other and to segment boundaries. The motivation is that if depressed speakers sound less animated, this should result in fewer peaks spaced more widely apart (a measure of speaking rate) and the peaks may be less extreme than for normal speakers. F0 is computed using default parameter settings for the snack PRAAT-style pitch tracker [39] and is used only for voiced regions according to the snack output. We expect pitch features to be robust to extrinsic variability, modulo the ability to detect pitch in a noisy or low-level signal. The f0-peak features record only the subset of pitch values found by an automatic peak-picking algorithm [40] run within each segment. Statistics computed in the f0peak-stats features include both pitch level and pitch peak distribution information. Pitch level includes the mean, max, and standard deviation of the peak pitches in the segment. Pitch peak distributions are intended to capture not pitch but rather the temporal distribution of pitch-accented syllables in the segment. These features include peak count, peak rate (count divided by segment duration), mean and maximum interpeak distances, and location of the maximum peak in the segment (e.g., early vs. late), each as a percentage of the way into the segment and as raw distance into the segment.

Intensity-related features, including *int* and *intpk*, are defined and computed in a manner similar to the pitch features, but for intensity rather than pitch. Intensity was computed using default intensity parameters in Praat [39]). Unlike pitch, we expect raw intensity values in int to reflect not only the speaker but also the recording session. Thus int and intpk are expected to partially reflect extrinsic factors. Table 2 provides a summary of all the feature types used in our experiments.

Table 2. Summary of all the features explored in this study

Name	Type	Extraction region	Feature dimension	Robustness to external session var.
DOCC	Acoustic	26 ms window at 10 ms frame rate	13	High
NMCC	Acoustic	26 ms window at 10 ms frame rate	13	High
MMeDuSA	Acoustic	52 ms window at 10 ms frame rate	16	High
GCC	Acoustic	26 ms window at 10 ms frame rate	13	Medium
AF	articulatory	20 ms window at 10 ms frame rate	8	High
AP	acoustic-phonetic	26 ms window at 10 ms frame rate	12	Medium
Tilt	vocal effort	voiced frames in segment	5	Medium
dle-on	vocal effort	voiceless->voiced transitions in segment	1	High
dle-off	vocal effort	voiced->voiceless transitions in segment	1	High
Encon	rhythmicity	200ms window in segment	7	High
Kaldi pitch	pitch	Frame	2	High
f0	pitch	Frame	1	High
f0pk	pitch at peaks	frames at peaks in segment	1	High
f0pk-stats	rhythmicity, rate, pitch	peak locations in segment	9 (stats)	High
Int	intensity	Frame	1	Low
Intpk	intensity at peaks	frames at peaks in segment	1	Low

In addition to the above features, we also used MFCC and IS13_ComParE features from the openSMILE toolkit [53] to benchmark the performance of the features explored in this work.

4. FEATURE MODELING AND MACHINE LEARNING SYSTEMS

The low-level spectral features (DOCC, NMCC, GCC and MMeDuSA) were time-contextualized using their delta features, and APs were contextualized using delta-delta features. The deltas and delta-deltas were calculated using the FILT approach detailed in [48], in which delta context was estimated with a window of 7 frames. The AFs were contextualized using shifted deltas, where deltas after a frame hop of two were contextualized using a delta spread of 3 and stacking 3 deltas from either side of the current frame. All frame-level features were mean- and variance-normalized on a per-subject basis. In our experiments, both frame-level and waveform-level features were used. We first converted all frame-level features to a waveform-level representation prior to training and classification. We found i-vectors [46] to be effective for this purpose. I-vectors are used in state-of-the-art speaker and language recognition technology [47, 46]. I-vectors use a generative modeling process, involving factor analysis on Baum-Welch statistics calculated from a Gaussian mixture model (GMM), to compress variable length feature vectors into a finite dimension vector. The i-vector *w* can be formulated as

$$s = m + Tw \qquad (2)$$

where *s* is the GMM mean supervector representing the Maximum A-Posteriori (MAP) point estimate of all data from a given recording, *m* the mean supervector from the Universal Background Model (UBM) trained on a large amount of AVEC data and *T* the low-rank i-vector subspace learned via factor analysis. Readers are directed to [46] for more details on the theory and implementation of i-vector subspace training and i-vector extraction. In contrast to the speaker and language recognition fields in which a large multitude of training data is readily available to train the i-vector subspace T, very limited training data were available for the AVEC challenge. To compensate, we restrained the dimension of the UBM to 16 Gaussian components and the i-vector subspace to only 30 dimensions. Variable length frame-level features extracted from the AVEC data were converted to 30-dimensional i-vectors using this framework. Note that the i-vector dimension was not optimized for each individual feature; our initial observations indicated that an i-vector dimension of 30 gave reasonable performance; thus, we kept that dimension fixed in all our subsequent experiments. Figure 8 illustrates this process. As an alternative to the i-vectors, we also explored GMM supervectors as fixed length representations of low-level (frame-based) features. From our experiments we always observed that the i-vectors perform better than the supervectors.

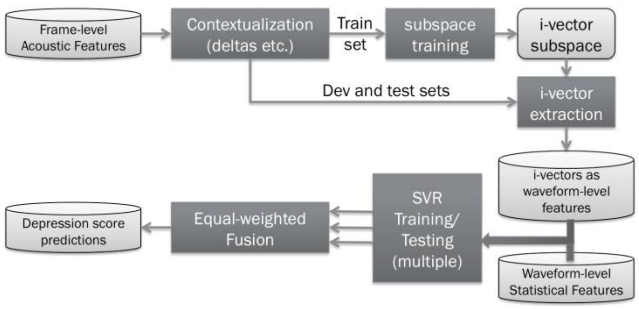

Figure 8. Depression prediction pipeline for individual features.

Once all recordings were represented as waveform-level vectors, we used support vector regression (SVR) [49] for classification. This was partially motivated by the success of SVR using i-vectors for the purpose of age estimation from speech [50]. In the context of AVEC, SVR provides a means of estimating depression scores between those evident in the SVR model training data (for instance in the higher end of the scale). The SVR training was performed using the *sklearn* python package [51] based on a polynomial kernel of order 20. All waveform-level vectors were first rank-normalized based on the training set. Rank normalization involves replacing each value of a vector with its ranked 'position' within the corresponding dimension of the training data. The rank values are scaled to the range [0-1]. Depression levels were then estimated from the development set. This process was conducted using multiple SVR classifiers using different feature combinations as illustrated in Figure 9.

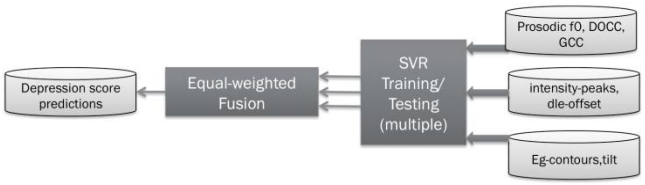

Figure 9. System combination

The depression scores from individual SVR classifiers were averaged to provide a means of equal-weight fusion. We used a held-out set of examples to train the fusion parameters. We observed that the benefit of fusion strategy failed to outweigh the loss in performance due to a reduction in system training data (after creating a held-out fusion set) for both i-vector extraction and SVR training.

We also investigated utterance-level features including intensity and contour features extracted from the audio stream. For these utterance level features, the i-vector approach was typically infeasible; instead, we used the average and, optionally, the standard deviation and maximum value. Furthermore, we noticed that for these features, the SVR classifier was suboptimal. We explored classifiers including neural networks (NN), a simple Gaussian mixture model (GMM) classifier and an 'extratrees' classifier as provided in the sklearn toolkit [51]. For the NN and extratrees classifiers, parameter tuning was performed using the development set for each feature. The simple Gaussian classifier was based on its use for audio characterization in [52].

5. EXPERIMENTAL RESULTS

The proposed architecture using both the frame-level and utterance-level features was evaluated on the audio part of the Depression Sub-Challenge (DSC). The evaluation criteria of depression detection were mean absolute error (MAE) and root mean square error (RMSE). Table 3 shows the error metrics for the development and test sets from the AVEC-2014 audio, video baseline systems [23] and our proposed system. Our final system was obtained from 6-way leave-one-out (LOO) score fusion of the 6 individual systems, where the 6 individual systems consisted of (1) i-vector fusion of MMeDuSA, AFs and NMCC features using SVR; (2) fusion of MFCC i-vectors, F0 i-vectors and F0-supervectors using SVR; (3) i-vector fusion of DOCC and GCC features using SVR; (4) fusion of KaldiF0 i-vectors, AP i-vectors and KaldiF0 supervectors using SVR; (5) dle-offset features using decision trees; and (6) intensity features using decision trees. The

result from this 6-way fusion on the development set gave results better than the AVEC-2014 DSC baseline for both audio- and video-only tasks. The fused system provided a relative reduction of 31.7% in MAE and 33.07% in RMSE compared to the 2014 audio-only DSC baseline system.

The lower half of Table 3 presents the results obtained from the 6-way fused system on the AVEC-2014 DSC audio-only test data. On the test data, our final system provided a relative reduction of 12.05% in MAE and 11.7% in RMSE compared to the 2014 audio-only baseline system. However, unlike the results on the development set, we did not observe a significant improvement in performance from our audio-only system compared to the video-only baseline DSC system. The 6-way fusion was optimized using the LOO score fusion as mentioned earlier using the development data. This may have given us a fusion configuration which may have failed to generalize well on the test data.

Table 4 shows the error metrics on the development data using the individual feature-based systems. From Table 4 we can observe that DOCCs provided both the lowest MAE and RMSE among all the low-level features. Note that the AVEC-2014 audio baseline feature consisted of 2268 dimensions, whereas the low-level features (after i-vector transformation) had only 30 dimensions. All the features in our experiments were used in their default configurations and we did not perform any optimization of feature configuration in any of our experiments. However, we explored different ways of contextualizing feature sets and observed that the use of delta-based contextualization helped across all the features. DOCC, GCC, NMCC and MMeDuSA features are known to be robust against noise corruption, and we have observed some degree of background noise in several AVEC-2014 audio files; this may have been the key reason why all of these features performed better than the MFCC features and were comparable in performance to the AVEC-2014 baseline.

For comparison purposes we also explored the Interspeech 2013 ComParE baseline features distributed with the openSMILE toolkit [53]. In addition to the full 6K-dimensional ComParE set, we also explored reducing the full set to a 47-dimensional set using the approach specified in [54]. The latter version yielded slightly lower error rates than the full set, but in both cases error rates were higher than for any of our single features (Table 4). Note that the ComParE feature sets are not expected to provide us results comparing to the baseline as they were not originally designed to for a depression level detection task.

Table 3. Performance of depression-level recognition for the development and test sets from the AVEC-2014 audio baseline systems and the SRI audio-only system. Baselines for the video only system are also shown in italics but cannot be directly compared.

	Features	MAE	RMSE
Dev. set	AVEC-2014 audio baseline	8.93	11.52
	AVEC-2014 video baseline	*7.58*	*9.31*
	SRI audio system	6.10	7.71
Test set	AVEC-2014 audio baseline	10.04	12.57
	AVEC-2014 video baseline	*8.86*	*10.86*
	SRI's audio system	8.83	11.10

From Table 3 we observe that baseline performance from video is better than from audio. We noticed impressionistically from listening to the data that annotations of the audio did not always match the overall annotation for depression. To better understand the role of audio features, we are currently performing a human annotation study in which the labeler has only the audio and no information on depression scores. We plan to rerun our classifiers using these additional labels and will report our findings to the community once completed.

Table 4. Performance by feature type (Note that none of the features were optimized for the given task)

Features	MAE	RMSE
DOCC	7.44	9.19
NMCC	8.14	9.92
MMeDuSA	8.01	10.05
GCC	7.53	9.36
AFs	8.52	10.78
APs	9.03	11.65
KaldiF0	8.99	11.43
Tilt	9.25	11.27
dle-on	9.23	10.77
dle-off	9.66	11.26
Encon	9.61	11.47
Kaldi Pitch	8.99	11.43
f0	9.99	13.74
f0pk	9.75	11.63
f0pk-stats	9.03	11.65
Int	10.16	12.00
Intpk	9.88	11.87
MFCC [53]	8.65	10.97

6. CONCLUSION

We have explored a wide array of features using the audio-only part of the AVEC- 2014 DSC sub-challenge. We have presented some features that have not previously been studied for this task. We have demonstrated that with suitable selection of low-dimensional features it is possible to outperform a baseline system that consists of feature dimensions of one order more. However, it is worth noting that the low-dimensional i-vector based features presented in this work are a result of sophisticated and complicated modeling strategies which the high-dimensional features lack. We observed that performance can be improved by performing fusion at two levels: (1) fusing features at the i-vector or supervector level, and (2) fusing scores from multiple systems. We did not explore tuning the feature parameters, such as the number of cepstral coefficients in the low-level features, normalization of those features, etc. In the future, we will explore the configuration of each feature individually to help us understand better how each of these features behaves for the task of depression level detection. We will also explore other modeling techniques and better fusion strategies to improve the performance of our system beyond what has been reported in this paper. Our

initial experiments have revealed that i-vector-level fusion of low-level features can result in more accurate systems. We are currently exploring feature-fusion approaches, where feature fusion will be explored not only among the low-level features, but also among frame- and utterance-level features.

7. ACKNOWLEDGMENTS

This articulatory features used in this research was supported by NSF Grant # IIS-1162046.

8. REFERENCES

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision, Washington, DC, American Psychiatric Association, 2000.

[2] M.M.Weissman, S. Wolk, R.B. Goldstein, D. Moreau, P. Adams, S. Greenwald, C.M. Klier, N.D. Ryan, R.E. Dahl, P. Wichramaratne, "Depressed adolescents grown up," Journal of the American Medical Association, 1999; 281(18):1701–1713.

[3] J. March, S. Silva, S. Petrycki, J. Curry, K. Wells, J. Fairbank, B. Burns, M. Domino, S. McNulty, B. Vitiello, J. Severe, "Treatment for Adolescents with Depression Study (TADS) team. Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents with Depression Study (TADS) randomized controlled trial," Journal of the American Medical Association, 2004; 292(7):807–820.

[4] J.A. Bridge, S. Iyengar, C.B. Salary, R.P. Barbe, B. Birmaher, H.A. Pincus, L. Ren, D.A. Brent, "Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment, a meta-analysis of randomized controlled trials," Journal of the American Medical Association, 2007; 297(15):1683–1696.

[5] J. Darby and H. Hollien, "Vocal and speech patterns of depressive patients," Folia phoniat, vol. 29, pp. 279–291, 1977.

[6] J. Darby, N. Simons, and P. Berger, "Speech and voice parameters of depression: A pilot study," J. Commun. Disorders , vol. 17, pp. 75–85, 1984.

[7] A. Ozdas, R. G. Shiavi, D. M. Wilkes, M. K. Silverman, and S. E. Silverman, "Analysis of vocal tract characteristics for near-term suicidal risk assessment," Methods of Information in Medicine, vol. 43, pp. 36–38, 2004.

[8] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," IEEE Transactions on Biomedical Engineering, vol. 51, no. 9, pp. 1530–1540, September 2004.

[9] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, "AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge," Proc. of AVEC 2013.

[10] L. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in IEEE Conference on Acoustics, Speech, and Signal Processing , Dallas, TX, USA, 2010, pp. 5154–5157.

[11] H. K. Keskinpala, T. Yingtha wornsuk, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, "Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands," in European Signal Processing Conference, Poznan, Poland, 2007, pp. 2229–2233.

[12] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, pp. 829–837, July 2000.

[13] E. M. II, M. A. Clements, J. W. Peifer, and L. Weisser, "Criticalanalysis of the impact of glottal features in the classification of clinical depression in speech," IEEE Transactions on Biomedical Engineering, vol. 55, no. 1, pp. 96–107, January 2008.

[14] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre, "Detecting depression from facial actions and vocal prosody," in International Conference on Affective Computing and Intelligent Interaction, 2009.

[15] T. Yingthawornsuk and R. G. Shiavi, "Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response," in International Conference on Control, Automation and Systems, Seoul, Korea, 2008, pp. 901–904.

[16] J. R. Williamson, R. Horwitz, T.F. Quatieri, B. Yu, B. S. Helfer, D. D. Mehta, "Vocal Biomarkers of Depression Based on Motor Incoordination," Proc. of AVEC 2013.

[17] N. Cummins, V. Sethu, J. Joshi, R. Goecke, A. Dhall, J. Epps "Diagnosis of Depression by Behavioural Signals: A Multimodal Approach," Proc. of AVEC 2013.

[18] H. Meng, H. Wang, H. Yang, M. Al-Shuraifi, Y. Wang, "Depression Recognition based on Dynamic Facial and Vocal Expression Features using Partial Least Square Regression," Proc. of AVEC 2013.

[19] B. Siddiquie, S. Khan, A. Divakaran, H. Sawhney "Affect Analysis in natural human interaction using joint hidden conditional random fields," Proc of ICME 2013.

[20] M. Amer, B. Siddiquie, S. Khan, A. Divakaran, H. Sawhney "Multimodal Fusion using Dynamic Hybrid Models", Proc. of WACV 2014.

[21] D. Maust, M. Cristancho, L. Gray, S. Rushing, C. Tjoa, and M. E. Thase, "Chapter 13 - Psychiatric rating scales," in Handbook of Clinical Neurology, vol. Volume 106, F. B. Michael J. Aminoff and F. S. Dick, Eds. Elsevier, 2012, pp. 227–237.

[22] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, W. Jarrold, "Using Prosodic and Spectral Features in Detecting Depression in Elderly Males", Proc. of Interspeech, 2011.

[23] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic "AVEC 2014 –3D Dimensional Affect and Depression Recognition Challenge," Proc. of AVEC2014

[24] A. Beck, R. Steer, R. Ball, and W. Ranieri, "Comparison of beck depression inventories -ia and -ii in psychiatric outpatients. Journal of Personality Assessment, 67(3):588{97, December 1996.

[25] V. Mitra, H. Franco, M. Graciarena, “Damped Oscillator Cepstral Coefficients for Robust Speech Recognition,” Proc. of Interspeech, pp. 886–890, 2013.

[26] V. Mitra, H. Franco, M. Graciarena, A. Mandal, “Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition,” Proc. of ICASSP, pp. 4117–4120, 2012.

[27] R. Drullman, J.M. Festen, R. Plomp, “Effect of Reducing Slow Temporal Modulations on Speech Reception,” J. Acoust. Soc. of Am., Vol. 95, No. 5, pp. 2670–2680, 1994.

[28] V. Ghitza, “On the Upper Cutoff Frequency of Auditory Critical-Band Envelope Detectors in the Context of Speech Perception,” J. Acoust. Soc. of America, vol. 110, no. 3, pp. 1628–1640, 2001.

[29] P. Maragos, J. Kaiser, T. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis,” IEEE Trans. Signal Processing, Vol. 41, pp. 3024–3051, 1993.

[30] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer and Y. Lei, “Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion”, in proc. of ICASSP 2013.

[31] A. Lawson, M. McLaren, Y. Lei, V. Mitra, N. Scheffer, L. Ferrer, M. Graciarena, "Improving Language Identification Robustness to Highly Channel-Degraded Speech Through Multiple System Fusion," in Proc. of Interspeech, pp. 1507-1510, Lyon, 2013.

[32] V. Mitra, M. McLaren, H. Franco, M. Graciarena, N. Scheffer, “Modulation Features for Noise Robust Speaker Identification,” Proc. of Interspeech, pp. 3703–3707, 2013.

[33] V. Mitra, H. Franco, M. Graciarena, D. Vergyri, “Medium duration modulation cepstral feature for robust speech recognition,” Proc. of ICASSP, pp. 1768-1772, Florence, 2014.

[34] H. Teager, “Some Observations on Oral Air Flow during Phonation,” IEEE Trans. ASSP, pp. 599–601, 1980.

[35] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, “Articulatory features from deep neural networks and their role in speech recognition,” Proc. of ICASSP, pp.3041-3045, Florence, 2014.

[36] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, L. Goldstein, “Articulatory Information for Noise Robust Speech Recognition,” IEEE Trans. on ASLP, Vol. 19, Iss. 7, pp. 1913–1924, 2010.

[37] H. Nam, L. Goldstein, E. Saltzman, D. Byrd, “TADA: An enhanced, Portable Task Dynamics Model in Matlab,” J. of Acoust. Soc. Am., 115(5), p. 2430, 2004.

[38] E. Shriberg, A. Stolcke, S. Ravuri, “Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style,” Proc. of Interspeech, 2013.

[39] P. Boersma, D. Weenink, “Praat: doing phonetics by computer,” Version 5.1.05, url: http://www.praat.org/, 2009

[40] N.C. Yoder, “Peak Finder,” Matlab program, url: http://www.mathworks.com/matlabcentral/fileexchange/25500-peakfinder, 2011.

[41] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpu "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in Proc. of ICASSP, 2014.

[42] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in Proc. ASRU, 2011.

[43] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", PhD thesis, University of Maryland College Park, December 2004.

[44] O. Deshmukh, J. Singh, C. Espy-Wilson. 2004. "A novel method for computation of periodicity, aperiodicity and pitch of speech signals," Proceedings of the 34th International Conference on Acoustics, Speech and Signal Processing, 17–21 May, Montreal, Canada, pp. 117–20.

[45] T. Pruthi, C. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," Proceedings of INTERSPEECH, pp. 1925–1928, 2007.

[46] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. on Speech and Audio Processing, 2011, 19, 788-798.

[47] D. Martınez, O. Plchot, L. Burget, O. Glembek, P. Matejka, “Language recognition in ivectors space.” Proceedings of Interspeech, Italy, 861-864, 2011.

[48] McLaren M.; Scheffer N.; Ferrer L. & Lei, Y. "Effective use of DCTs for Contextualizing Features for Speaker Recognition," Proc. ICASSP, 2014.

[49] H. Drucker, C.J. Burges, L. Kaufman, A. Smola, V. Vapnik, "Support vector regression machines. Advances in neural information processing systems," 9, pp. 155-161, 1997

[50] M. H. Bahari, M. McLaren, H. van hamme, and D. A. van Leeuwen. "Age estimation from telephone speech using i-vectors," in Proc. of InterSpeech 2012, 2012.

[51] Pedregosa et al. "Scikit-learn: Machine Learning in Python," JMLR 12, pp. 2825-2830, 2011. url: http://scikit-learn.org

[52] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, “A unified approach for audio characterization and its application to speaker recognition,” in Proc. of the Speaker and Language Recognition Workshop, Odyssey 2010, Brno, Czech Republic, Jun. 2010.

[53] F. Eyben, M. Wöllmer, B. Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10.2010.

[54] K. Subrahmanyam, N. Shiva Sankar, S. Praveen Baggam, R. Rao S, "A Modified KS - test for Feature Selection," IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2278-8727, Vol. 13, Iss. 3, pp. 73-79, 2013.