Published in final edited form as:

IEEE Int Workshop Mach Learn Signal Process. 2014 September ; 2014: . doi:10.1109/mlsp.2014.6958856.

INFERRING CLINICAL DEPRESSION FROM SPEECH AND SPOKEN UTTERANCES

Meysam Asgari,
Center for Spoken Language Understanding Oregon Health & Science University, Portland, Oregon

Izhak Shafran,
Center for Spoken Language Understanding Oregon Health & Science University, Portland, Oregon

Lisa B. Sheeber
Oregon Research Institute Eugene, Oregon

Abstract

In this paper, we investigate the problem of detecting depression from recordings of subjects' speech using speech processing and machine learning. There has been considerable interest in this problem in recent years due to the potential for developing objective assessments from real-world behaviors, which may provide valuable supplementary clinical information or may be useful in screening. The cues for depression may be present in ”what is said” (content) and ”how it is said” (prosody). Given the limited amounts of text data, even in this relatively large study, it is difficult to employ standard method of learning models from n-gram features. Instead, we learn models using word representations in an alternative feature space of valence and arousal. This is akin to embedding words into a real vector space albeit with manual ratings instead of those learned with deep neural networks [1]. For extracting prosody, we employ standard feature extractors such as those implemented in *openSMILE* and compare them with features extracted from harmonic models that we have been developing in recent years. Our experiments show that our features from harmonic model improve the performance of detecting depression from spoken utterances than other alternatives. The context features provide additional improvements to achieve an accuracy of about 74%, sufficient to be useful in screening applications.

Keywords

Depression; Speech analysis; Telemedicine

1. INTRODUCTION

Clinical depression is a common disorder that negatively affects person's health, mood, thoughts, behavior, work, family, and ability to function in everyday life [2] [3]. It is often undiagnosed and afflicts a large portion of the population, for example, over 19 million American adults [4]. The diagnosis is subjectively performed by an expert practitioner based on patient's mental state driven from interviews and self-report experiences. This assessment is costly, time-consuming, and often requires patients' presence at the clinic. Recent studies have explored the influence of emotional changes on phonatory and articulatory of speech production system [5] [6]. These observations have motivated researches to explore alternative approaches based on speech processing techniques, which can be used in real applications such as automatically screening and telemonitoring of depressive disorders. Subsequently, a number of studies have attempted to find potential clues in subjects' speech that reflect influences of mood disorders [7]. Acoustic features of speech signal including pitch, formants, Harmonic-to-Noise Ratio (HNR), shimmer, jitter, speech rate, energy, and glottal features have been used to analyze voices of patients with depression [8] [9] [10].

Speech pathologists characterize speech from clinical depressed patients' as monotone, mono-loud, and lifeless [11]. There has been a considerable interest on analyzing acoustic properties of speech for the hope of quantitative assessment of clinical depression [12] [13] [14]. Moore and his colleagues [15] attempted to employ prosodic (pitch, energy, speech rate), vocal tract (first, second, and third formant frequencies and their bandwidths), and glottal features (starting points of glottal opening and closing, minimum point in glottal derivative, maximum glottal opening) for classifying 15 depressed from 18 control (15 males and 18 females) subjects. They achieved a classification accuracy of 96% in a leave-one-out cross validation strategy by combination of vocal and glottal features. This study was conducted on a relatively small data set and its results may not generalize. Recently, Low and his colleagues introduced an automatic approach to classify 68 clinically depressed adolescents from 71 controls. They employed a variety of features – spectral, prosodic, cepstral, glottal and features derived from a non-linear operation, Teager energy operator (TEO). The performance of different combination of features by GMM and SVM classifiers show classification accuracy of 81%-87% for males and 72%-79% for females. In another study, Algowinem and his colleagues [16] applied several machine learning strategies including hierarchal fuzzy signature (HFS) and multi-layer perceptron (MLP) classifiers on a broad range of speech measures. They found that loudness, intensity, and root mean square were the most useful features.

2. CORPUS

Our corpus for this study was collected by Oregon research institute (ORI) and consists of video recordings of adolescents subjects' during their interaction with their family. Subjects were asked to participate in three different 20-minutes interactions with their parents: event-planning interaction (EPI), problem-solving interaction (PSI), family consensus interaction (FCI) [17]. All interaction were administrated by a trained interviewer in a quiet room at ORI. The recordings were collected from 148 subjects, including 98 females and 50 males, 14 to 18 year old. Of these subjects, based on clinical assessment, 71 adolescents (50 females and 21 males) were diagnosed depressed and 77 individuals (48 females and 29 males) were healthy controls.

As a clinical reference, the severity of subject's condition were coded by living-in-family-environment (LIFE) coding system [17]. The LIFE coding system was developed for assessment of behavioral characteristic of individuals with depressive disorders. In LIFE coding system, behavior is coded based on verbal, nonverbal, and para-verbal behavior and codes do not necessarily imply whether subject is speaking. A group of psychologists at ORI coded subjects’ verbal content and emotional state in terms of 27 contents code and 10 affect code available on the LIFE coding system. Subjects’ behavior were marked by six categories of *angry*, *dysphoric*, *happy*, *neutral*, *end*, and *other* per each second in family interaction sessions. For the purpose of our experiments, we extracted audio from the video recordings and converted them from stereo to mono channel format. Then, speech segments annotated with *angry*, *dysphoric*, and *happy* tags where chopped and concatenated for the manual transcription. There was noise in the annotations as they were not always aligned with the utterances. We asked annotators to manually identify the speaker at each segment to alleviate noise in the labels. Our experiments were performed on segments from the adolescent subjects.

3. SPEECH FEATURES

In our experiment, we used a broad range of features to capture speech clues associated with clinical depression. For our baseline system, we adopted the baseline features defined in INTERSPEECH 2010 Paralinguistic Challenge [18] using *openSMILE* toolkit [19].

The features, comprised of 1582 components, can be broadly categorized into three groups: 1) loudness related features such as RMS energy and PCM loudness, 2) voicing related features like pitch frequency, jitter, and shimmer., and 3) articulatory related features such as mel-frequency cepstral coefficients and line spectral frequencies. The features computed at the frame-level were summarized into a global feature vector of fixed dimension for each recording using 21 standard statistical functions including min, max, mean, skewness, quartiles and percentile.

4. TEXTUAL FEATURES

In order to gauge the effect of speech contents in the clinical depression, we extracted textual features from manual transcripts. To extract features from text, we used a published table of valence and arousal ratings by Warriner et al. [20] to tag each word in an utterance with an arousal and a valence rating. This is akin to projecting the words into an embedding space, albeit not learned from data as in the recent deep neural network literature [1]. We computed the per-utterance mean, standard deviation, minimum, and maximum. For missing words we imputed valence and arousal by randomly drawing 5 words from the table and computing their average.

5. SPEECH FEATURES FROM HARMONIC MODEL

Alternatively, we extracted prosodic features using the harmonic model [21] [22] [23] [24]. In previous work, we found that these features detect voiced segments and estimate pitch frequency more accurately than other algorithms and that they are useful in rating the severity of subjects’ Parkinsons disease [25].

5.1. Harmonic Model

Briefly, in the harmonic model, the speech samples are represented as follows. Let $\mathbf{y} = [y(t_1), y(t_2), \dots, y(t_N)]^T$ denote the speech samples in a voiced frame, measured at times $t_1, t_2, \dots,$

t_T. The samples can be represented by the harmonic model with additive noise $\mathbf{n} = [n(t_1), n(t_2), \dots, n(t_N)]^T$ as follows:

$$\begin{aligned} s(t) &= a_0 + \sum_{h=1}^{H} a_h cos(2\pi f_0 ht) + b_h sin(2\pi f_0 ht) \\ y(t) &= s(t) + n(t) \end{aligned} \tag{1}$$

where H denotes the number of harmonics and $2\pi f_0$ stands for the fundamental angular frequency. Assuming the noise distribution is constant during the frame and is given by $\mathcal{N}(0, \sigma_n^2)$, estimation of the unknown parameters $\Theta = [f_0, a_h, b_h, \sigma_n^2, H]$ can be cast into a maximum likelihood (ML) framework [23]. However, ML estimation of the pitch period may lead to pitch halving and doubling errors. We addressed this problem in our previous work and improved the robustness of the pitch estimates by smoothing the likelihood function [26]. Given the estimates of Θ, we can reconstruct the speech signal for the further analysis.

5.1.1. Vector of Harmonic Coefficients—We showed in our earlier work that estimating the number of harmonics, a model order selection problem, can be reliably solved using a Bayesian information criteria (BIC) [25]. Given the number of harmonics, H, we estimate the coefficient of harmonics, a $(2H + 1)$-dimensional feature vector, $[a_0, a_1, \dots, a_H, b_1, \dots, b_H]^T$. We then transform the feature vector to the log-domain after taking its absolute value.

5.2. Jitter and Shimmer

Jitter and shimmer refer to a short-term (cycle-to-cycle) perturbation in the pitch frequency and the amplitude of voice waveform respectively. Perturbation analysis is based on the fact that small fluctuations in frequency, and amplitude of waveform reflect the inherent noise of voice. These measures can be sensitive to noise. Traditionally, computation of jitter and shimmer assumes that these parameters are constant during the frame. Alternatively, estimate jitter and shimmer from the signal reconstructed using the estimated parameters of the harmonic model.

5.2.1. Shimmer—Speech can be considered as an amplitude modulated (AM) and frequency modulated (FM) signal. To compute shimmer, we first represent the speech waveform using the harmonic model with time-varying amplitudes (HM-VA) [22] as shown in Equation 2.

$$\begin{aligned} s(t) &= a_0(t) + \sum_{h=1}^{H} [a_h(t) \cos(2\pi f_0 ht)] \\ &+ \sum_{h=1}^{H} [b_h(t) \sin(2\pi f_0 ht)] \end{aligned} \tag{2}$$

Note, this is different from the harmonic model represented previously in Equation 1. Unlike, the previous model whose harmonic coefficients are fixed, in the time-varying model, as the name implies, the coefficients are allowed to vary $a_h(t)$ and $b_h(t)$ over time. Thus, this model is capable of capturing sample to sample variation in harmonic amplitude within a frame. Continuity constraints can be imposed using a small number of basis functions ψ_i as in Equation 3 [22].

$$a_h(t) \;=\; \sum_{i=1}^{I} \alpha_{i,h}\psi_i(t)\,, \quad b_h(t) \;=\; \sum_{i=1}^{I} \beta_{i,h}\psi_i(t) \tag{3}$$

In our experiments, we use 4 Hanning windows, ($I = 4$), within a frame, centered at 0, $M/3$, $2M/3$, and M. Each basis function is $2M/3$ samples long and has an overlap of $M/3$ with immediate adjacent window. Given the estimate fundamental frequency from Equation 1, we compute $a_h(t)$ and $b_h(t)$ using a maximum likelihood framework. Shimmer can be considered as a temporal function, $f(t)$, that scales the amplitudes of all the harmonics in the time-varying model.

$$c_h(t) = c_h f(t) + e(t), \quad t = 1, \ldots, T, h = 1, \ldots, H \tag{4}$$

where $c_h = \sqrt{\Sigma_{h=1}^{H}\, a_h^2 + b_h^2}$ denotes the amplitude of the harmonic components in harmonic model with constant amplitudes and $c_h(t)$ is the counterpart from the time-varying model. Once again, assuming uncorrelated noise, f(t) can be estimated using maximum likelihood criterion.

$$\hat{f}(t) = \frac{\Sigma_{h=1}^{H}\, c_h c_h(t)}{\Sigma_{h=1}^{H}\, c_h^2} \tag{5}$$

The larger the tremor in voice, the larger the variation in $f(t)$. Hence, we use the standard deviation of $f(t)$ as a summary statistics to quantify the shimmer. Figure 1 illustrates an example frame, the signal estimated using the harmonic model with constant amplitude and with time-varying amplitudes. The signal estimated with the time-varying harmonic amplitudes is able to follow variations not only in amplitude but also variation in pitch to a certain extent. Also, dotted red line in this figure shows the the envelop of speech waveform extracted for computing the shimmer.

5.2.2. Jitter—To estimate jitter, we create a matched filter using a one pitch period long segment from the reconstructed signal and convolve it with the original speech waveform [27]. The distance between the maxima in the convolved signal defines the pitch periods. The perturbation of the period is normalized with respect to the given pitch period and its standard deviation is an estimate of the jitter. Thus, this method allows the computation of jitter within the 25ms analysis window.

5.3. Harmonic to noise ratio (HNR)

Once the parameters of the harmonic model are estimated for a frame, the noise can be computed by subtracting the reconstructed signal from the original speech signal. Given the estimated HM parameters for each frame, the HNR and the ratio of the energy in the first and the second harmonics (H12) can be computed as follows.

$$
\begin{aligned}
c_h &= \sqrt{\sum_{h=1}^{H} a_h^2 + b_h^2} \\
HNR &= \log \sum_{h=1}^{H} c_h^2 - \log \sum_{t=1}^{N} \left(y(t) - s(t)\right)^2 \\
H12 &= \log c_1 - \log c_2
\end{aligned}
\qquad (6)
$$

5.4. Per-Utterance Feature Vectors

We extract 25 ms long frames using a Hanning window with a 10 ms shift. We first detect voiced frames robustly by calculating the likelihood of voicing under the harmonic model. The voicing decision at the segment level is computed by formulating a one-state hidden Markov model (HMM) [28]. The state could either be voiced or unvoiced, with likelihood given by the per-frame harmonic model. The transition model consists of a simple zero-mean Gaussian. We compute voicing decision over the utterance using Viterbi alignment. Subsequently, we compute various voicing related features for voiced frames, including pitch frequency, HNR, H12, jitter, shimmer, and harmonic coefficients. These pitch-related features are combined with standard features including energy, spectral entropy, and MFCCs. For unvoiced frames, we just compute energy, spectral entropy, and MFCCs features. Features extracted from voiced regions tend to differ in nature compared to those from unvoiced regions. These differences were preserved and features were summarized in voiced and unvoiced regions separately. Per-utterance features are computed by applying standard summary statistics such as mean, median, variance, minimum and maximum to the per-frame voiced (unvoiced) features, generating a 192-dimensional per-utterance feature vector.

5.5. Experiments

We compared the performance of a SVM classifier on our data using 30-fold cross-validation for classifying depressed from control subjects. Table 1 reports the performance of classifiers trained on different feature sets. The SVM classifier with several kernel functions including linear, polynomial, and radial basis function were employed from open-source Scikit-learn toolkit [29]. Parameters of the optimal classifier were determined via grid search and cross validation on training set. The best performance among all models was obtained with the linear kernel with the exception of the model with *openSMILE* feature set, for which RBF kernel was better. Speech features, extracted from both *openSMILE* and harmonic model, perform significantly better than chance with a p-value of less than 0.01, according to cross-validated paired t-test [30] and denoted by (†) in the table. Also, Table 1 indicates that incorporating textual features result in an additional improvement. However, solely use of textual features didn’t yield in a significant performance compare to chance. This can be justified by the fact that word-based approach for feature extraction is simplistic and it’s not capable of extracting contextual dependencies.

5.5.1. Effectiveness of family interaction sessions—In this section, we separately examine the influence of each family interaction session in classifying the depressed speech. As we mentioned earlier, there are three types of family interactions: 1) EPI, 2) PSI, and 3) FCI. These tasks differently evoke emotional state of adolescents during the family interaction and potentially reveal different aspects of their behavior. For instance, PSI tends to elicit the conflictual behavior of adolescents when interact with their parents. Table 2 reports the performance of SVM classifiers The results indicate that features extracted from PSI are most effective in this task. The performance of classifier trained on features extracted from harmonic model is significantly better than chance with p-value of less than 0.01.

6. CONCLUSIONS

This paper investigates the problem of detecting depression from recordings of subjects' spoken utterances. Given the scarcity of the text data for training models with n-grams, we explore an alternative method to extract content information related to affect by encoding words in terms of valence and arousal, using a look up table that has been compiled by averaging responses from large number of raters. We extract novel acoustic and prosodic features from harmonic models and find that they outperform standard features such as those computed from *openSMILE*. The textual features provide additional gain, achieving a classification accuracy of about 74%.

ACKNOWLEDGMENT

This research was supported by Google, Intel and IBM awards as well as NSF awards IIS 0964102 and 1027834, and NIH award K25 AG033723. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the funding agencies.

8. REFERENCES

[1]. Weston J, Ratle F, Mobahi H, and Collobert R, "Deep learning via semi-supervised embedding," in Neural Networks: Tricks of the Trade. Springer, 2012, pp. 639–655.

[2]. Pine DS, Cohen E, Cohen P, and Brook J, "Adolescent depressive symptoms as predictors of adult depression: moodiness or mood disorder?" American Journal of Psychiatry, vol. 156, no. 1, pp. 133–135, 1999. [PubMed: 9892310]

[3]. Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A et al., "Influence of life stress on depression: moderation by a polymorphism in the 5-htt gene," Science Signaling, vol. 301, no. 5631, p. 386, 2003.

[4]. Kessler RC, Chiu WT, Demler O, and Walters EE, "Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication," Archives of general psychiatry, vol. 62, no. 6, p. 617, 2005. [PubMed: 15939839]

[5]. Moses PJ, "The voice of neurosis." 1954.

[6]. Asgari M, Kiss G, van Santen J, Shafran I, and Song X, "Automatic measurement of affective valence and arousal in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on IEEE, 2014, pp. 965–969.

[7]. Alpert M, Pouget ER, and Silva RR, "Reflections of depression in acoustic measures of the patients speech," Journal of affective disorders, vol. 66, no. 1, pp. 59–69, 2001. [PubMed: 11532533]

[8]. Hargreaves WA and Starkweather JA, "Voice quality changes in depression," Language and Speech, vol. 7, no. 2, pp. 84–88, 1964.

[9]. Nilsonne Å, Sundberg J, Ternstrom S, and Askenfelt A, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," The Journal of the Acoustical Society of America, vol. 83, p. 716, 1988. [PubMed: 3351130]

[10]. Tolkmitt F, Helfrich H, Standke R, and Scherer KR, "Vocal indicators of psychiatric treatment effects in depressives and schizophrenics," Journal of communication disorders, vol. 15, no. 3, pp. 209–222, 1982. [PubMed: 7096618]

[11]. Scherer KR, "Expression of emotion in voice and music," Journal of voice, vol. 9, no. 3, pp. 235–248, 1995. [PubMed: 8541967]

[12]. Beck AT, Ward CH, Mendelson M, Mock J, and Erbaugh J, "An inventory for measuring depression," Archives of general psychiatry, vol. 4, no. 6, p. 561, 1961. [PubMed: 13688369]

[13]. Low L-SA, Maddage NC, Lech M, Sheeber L, and Allen N, "Content based clinical depression detection in adolescents," 17th EUSIPCO, pp. 24–28, 2009.

[14]. Low L-S, Maddage M, Lech M, Sheeber LB, and Allen NB, "Detection of clinical depression in adolescents speech during family interactions," Biomedical Engineering, IEEE Transactions on, vol. 58, no. 3, pp. 574–586, 2011.

[15]. Moore E, Clements MA, Peifer JW, and Weisser L, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," Biomedical Engineering, IEEE Transactions on, vol. 55, no. 1, pp. 96–107, 2008.

[16]. Alghowinem S, Goecke R, Wagner M, Epps J, Gedeon T, Breakspear M, and Parker G, "A comparative study of different classifiers for detecting depression from spontaneous speech."

[17]. Hops H, Living in Family Environments (LIFE) coding system: Reference manual for coders. Oregon Research Institute, 1995.

[18]. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller CA, and Narayanan SS, "The interspeech 2010 paralinguistic challenge." in INTERSPEECH, 2010, pp. 2794–2797.

[19]. Eyben F, Wöllmer M, and Schuller B, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the international conference on Multimedia ACM, 2010, pp. 1459–1462.

[20]. Warriner AB, Kuperman V, and Brysbaert M, "Norms of valence, arousal, and dominance for 13,915 english lemmas," Behavior research methods, vol. 45, no. 4, pp. 1191–1207, 2013. [PubMed: 23404613]

[21]. Stylianou Y, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modifycation," in Ph.D. dissertation, Ecole Nationale des Tlcomminications, 1996.

[22]. Godsill S and Davy M, "Bayesian harmonic models for musical pitch estimation and analysis," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, 2002, pp. 1769–72.

[23]. Tabrikian J, Dubnov S, and Dickalov Y, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," IEEE Transactions on Speech and Audio Processing, vol. 12, no. 1, pp. 76–87, 2004.

[24]. Christensen MG and Jakobsson A, "Multi-pitch estimation," Synthesis Lectures on Speech & Audio Processing, vol. 5, no. 1, pp. 1–160, 2009.

[25]. Bayestehtashk A, Asgari M, Shafran I, and McNames J, "Fully automated assessment of the severity of parkinson's disease from speech," Computer Speech & Language, 2013.

[26]. Asgari M, Bayestehtashk A, and Shafran I, "Robust and accurate features for detecting and diagnosing autism spectrum disorders."

[27]. Asgari M and Shafran I, "Extracting cues from speech for predicting severity of parkinson's disease," in Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on IEEE, 2010, pp. 462–467.

[28]. Asgari M, Shafran I, and Bayestehtashk A, "Robust detection of voiced segments in samples of everyday conversations using unsupervised hmms," in Spoken Language Technology Workshop (SLT), 2012 IEEE IEEE, 2012, pp. 438–442.

[29]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al., "Scikit-learn: Machine learning in python," The Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[30]. Dietterich TG, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural computation, vol. 10, no. 7, pp. 1895–1923, 1998. [PubMed: 9744903]

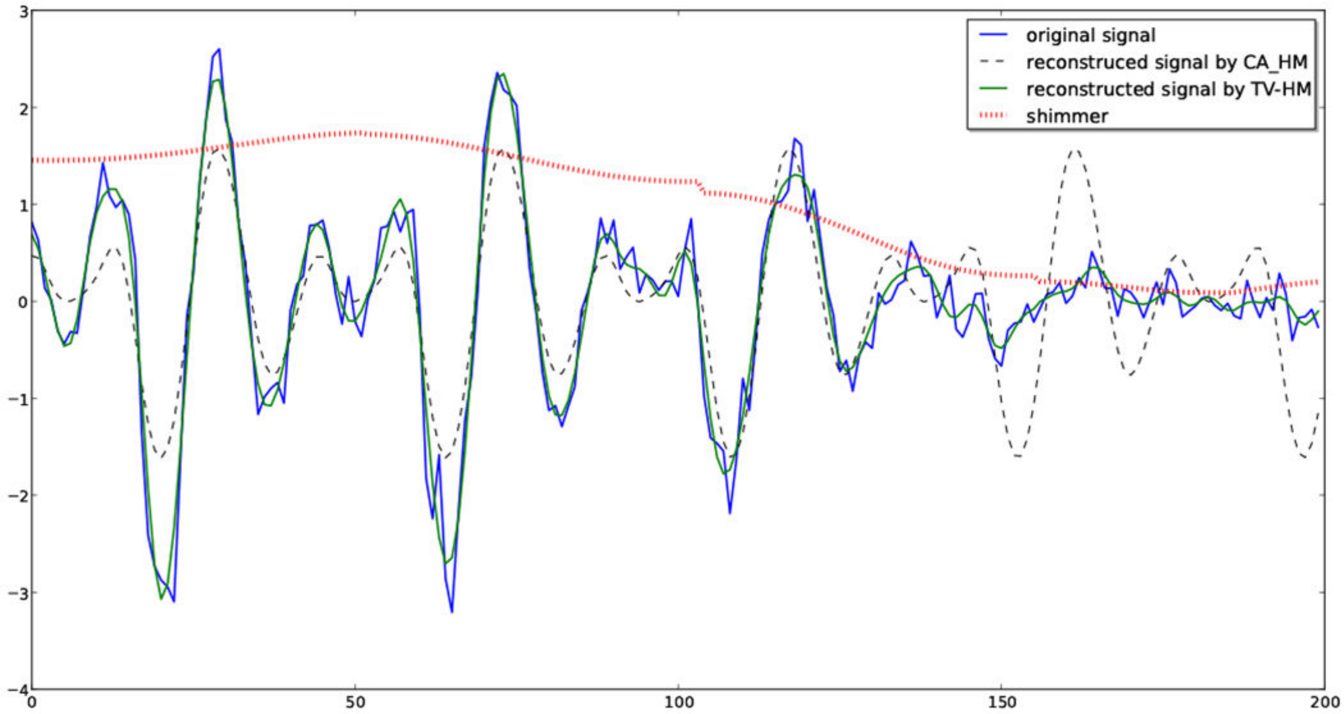

Fig. 1.

An example speech frame (blue), estimated signal from harmonic model with time-varying amplitude (green), estimated signal from harmonic model with constant amplitude (black), and estimated shimmer (red).

Table 1.

Comparison of performance of SVM classifier using different features for classifying clinical depression of adolesents.

Features	**Classification Accuracy**
Chance	52.4
Text	65.4
openSMILE	64.7†
openSMILE+Text	68.0†
Harmonic Model	68.7†
Harmonic Model+Text	74.0†

Table 2.

Effect of family interaction sessions on classifying clinical depression in adolesents.

Features	EPI	PSI	FCI
Chance	49.2	49.2	49.2
openSMILE	60.0	64.7	56.0
Harmonic Model	66.1	71.4†	57.6