Depression Prediction Via Acoustic Analysis of Formulaic Word Fillers

Brian Stasak[1,2], Julien Epps[1,2] and Nicholas Cummins[3]

[1] School of Elec. Eng. & Telecomm., University of New South Wales, Sydney, Australia
[2] Data61-CSIRO, Sydney, Australia
[3] Chair of Complex and Intelligent Systems, University of Passau, Passau, Germany

brian.stasak@student.unsw.edu.au, j.epps@unsw.edu.au, nicholas.cummins@uni-passau.de

Abstract

Understanding what kind of speech is most effective for predicting depression deserves more attention since diagnosis and monitoring is often based on limited interview or questionnaire responses. Consequently, this paper investigates thin slice data selection using token word analysis, demonstrating that spoken formulaic language and in particular filler words hold acoustic discriminative properties that are useful when predicting different ranges of depression. Results from an analysis of the DAIC corpus indicate that filler words are equally or more effective for depression prediction as using entire utterances, and that acoustic and linguistic features combined generate competitive depression prediction results.

Index Terms: depression, formulaic language, speech emotion, thin slice.

1. Introduction

Continuous digital bio-signal analysis is widely used in many fields of healthcare, such as neurology, cardiology, and physiology. In recent years, speech has gained attention as a bio-signal measure aiding mental health diagnoses and monitoring. Automated speech processing, although not fully implemented as a standard protocol tool in assessing and monitoring depressed individuals, is not far off in the future as a potential clinical application. The measurement and analysis of vocal characteristics is advantageous over other bio-signals because of its naturalistic communicative form and non-invasive collection without complex expensive machines that require specialized training (e.g. fMRI, PET, SQUID).

Notwithstanding the ease of collecting speech recordings, the dilemma of knowing which vocal segments carry compact informational value is a questioned predicament by researchers. The concept of 'thin slice' data selection was studied in [1], wherein brief social or clinical observations can often yield useful compact information rather than longer observations. Researchers surmise several considerations regarding thin slice data selection theory: (1) the channel of observations, whether verbal or non-verbal, has little effect on the predictive observational results; (2) a great deal of behavioral affect is generated unintentionally or unconsciously yet it still contributes to other people's observed predictions or interpretations; (3) when affective thin slice data selection is proven effective it can significantly reduce resources without sacrificing performance; and (4) while examining human behavior the thin slice approach works particularly well when predicting vital interpersonally oriented criterion variables [1].

The concept of data selection has proven applicable in many areas of speech processing. Prior research indicates in speech processing applications, such as speech recognition, speaker identification, and speech emotion classification, data selection that reduces phonetic variability and compares similar phonetic structures can improve performance. For instance, Reynolds et al. [2] demonstrated by focusing on specific acoustic classes (i.e. text-dependent words) individual speaker models develop better modeling of short-term variations thereby resulting in higher overall identification performance for short utterances. In Boakye et al. [3], researchers analyzed thirteen token words consisting of less than ten-percent of all utterances across speakers. Their results showed speaker identification performance was superior for short fixed token words than for an entire set of words.

Researchers have also advocated for the inclusion of linguistic features with acoustic features in paralinguistic applications. For instance, Shriberg and Stolcke [4] and Ishihara et al. [5] analyzed text transcripts to generate token word linguistic features that contributed to higher speaker recognition performance. Research has also shown habitual speaker idiolects and formulaic language use can provide unique speaker information [3]. Formulaic language is common in everyday discourse and even monolog narratives, where it contributes to nearly a quarter of all conversational speech [6]. It includes conventional word expressions, proverbs, idioms, expletives, hedges, bundles, and fillers.

In brief, depressed speaker characteristics reported to date include: reduction in vocal prosody dynamics; decreased vocal pitch; reduced vocal intensity; slower rate of speech; increased vocal tenseness; motor incoordination, including disfluency or motor retardation [7]; reduced vowel space ranges [8]; and difficulty with word retrieval [9]. Further, investigation into spoken phonetic variability across individual phonemes has indicated potential biomarkers for depression disorders which are based on speech rate and phoneme durations [10].

In the literature, only a limited number of *non-acoustic* linguistic text-based studies have specifically examined formulaic word fillers in depressed populations [6, 11]. The *acoustic* evaluation of formulaic language for predicting levels of depression has several practical advantages. Its related filler words are found spontaneously occurring in large numbers across a wide range of speakers irrespective of gender, age, language, and education. Furthermore, the constrained acoustic phonetic variability that arises naturally from considering only token filler words facilitates intra- and inter-speaker comparison. *Intra-speaker*, there are typically multiple examples of each token word per utterance, helping to construct a more focused analysis of phoneme characteristics. *Inter-speaker*, token word characteristics can also be more readily compared between speakers than entire utterances. Studies have demonstrated that even a single phoneme type in large quantities can help to reveal discriminative information regarding a speaker's emotional state [8, 12].

While formulaic language has arguably only a minor contribution to semantic/pragmatic content, it adds importantly to cognitive-emotive speaker internalization [6]. In this paper, it is hypothesized that since formulaic fillers are quite common in number and represent speakers' mental internalization, they will reveal more about the effects of depression on speech than utterances in general.

2. Speech Corpus

The audio portion of the training and development from the Distress Analysis Interview Corpus (DAIC) [9][1] was used for all experiments herein. The DAIC was designed to investigate language, nonverbal behaviors, psychophysiology, and assisted human-computer commutative dialog. This database was chosen because it provides a fixed set of utterances which were spoken by a computer generated interviewer; a large group of speakers, 143 males and females; high-quality close-talking microphone recordings; natural speech in a clinical type environment; and PHQ-8 evaluations along with scores per participant. The PHQ-8 is a popular eight-questioned self-administered mental health assessment tool commonly used in diagnosis of depression disorders [13]. It has an interval scale of 0 to 24, where larger scores imply a greater depression severity. The DAIC was also chosen because it includes phrase-level transcripts with beginning/ending time markers, which made extracting single token words possible with minimal error. Only individually segmented token word entries were evaluated. These segments were determined based on a transcriber's indication of when single word tokens began and ended. For more information regarding the DAIC transcription conventions see [9]. The token words evaluated in all experiments are listed in Table 1. The ten token words combined included 95% of speakers from the train and 100% of speakers from the development sets; the other 5% were omitted due to transcript time marker errors.

Table 1: *Description of token words evaluated, percentage of training/test speaker coverage, total number of utterances, and number of unique speakers in the DAIC.*

Token Words	Word Type	% Train	% Test	# Total	# of Unique Speakers
"Hmm"	Filler	33%	49%	139	52
"Mhm"	Filler	35%	43%	123	53
"Mm"	Filler	48%	60%	168	72
"Uh"	Filler	52%	60%	298	77
"Umm"	Filler	85%	94%	1276	124
"So"	Filler	27%	54%	119	48
"No"	Polar	77%	74%	230	109
"Yeah"	Polar	60%	66%	305	88
"Okay"	Polar	27%	43%	56	44
"You Know"	Bundle	21%	23%	57	31

3. Experimental Methodology

3.1. Feature Overview

The baseline experiments used the 88 eGeMAPS acoustic features [14]. Additionally, there were 116 acoustic features extracted using VoiceSauce [15]. Examples of the eGeMAPS and VoiceSauce functional features include median and standard deviations for jitter, shimmer, Mel-Frequency Cepstral Coefficients (MFCC), pitch, formant frequencies, formant bandwidths, formant amplitudes, and harmonic-to-noise ratios. For all acoustic features, windows of 20 ms (with 10 ms overlap) were applied.

Linguistic features were derived from individual speaker's entire recording session transcript and compiled using text-processing scripts. Although many linguistic features were considered, the average utterance length, average syllables per second, percentage of unique words, percentage of articles/prepositions/pronouns, and readability scores per speaker were most valuable. Readability scores were based on common methods, such as Flesh-Kincaid Grade Level and Gunning Fog Index [16]. While these readability scores are typically derived from written passages, they can also be useful when qualitatively applied to verbal transcripts.

3.2. System Design

The AVEC 2016 depression prediction sub-challenge baseline [17] utilised all training and development data in the DAIC, applied similar acoustic features, and employed a support vector machine for regression analysis. The baseline acoustic features were created using entire utterances, and a depression prediction baseline of 5.35 Mean Absolute Error (MAE) and 6.74 Root-Mean Squared Error (RMSE) was achieved. For comparison with this baseline, Support Vector Regression (SVR) was also used to predict the depression scores for experiments herein. SVR has been successfully applied to speech depression/emotion prediction tasks and is known for effective statistical generalization [18]. Based on the SVR output, two standard performance metrics were used to evaluate the overall predictive accuracy due to their application in recent speech depression prediction challenges: MAE and RMSE. One distinct advantage the RMSE has over the MAE is it does not utilize absolute values and is generally better at revealing model performance differences.

In Figure 1, the system design involves two main inputs, speech and spoken transcript data. During the acoustic and linguistic feature extraction stage, feature selection can be performed (indicated by shaded boxes). The feature selection process retains the most salient features and omits any weaker features for statistical predictive analysis. Note that occasionally some speakers' token words were too short in verbal duration, so some acoustic features could not be adequately calculated (producing nulls) and/or some linguistic features were found to have little variance. Afterwards, a depression score prediction output was generated and compared to the ground truth depression PHQ-8 score.

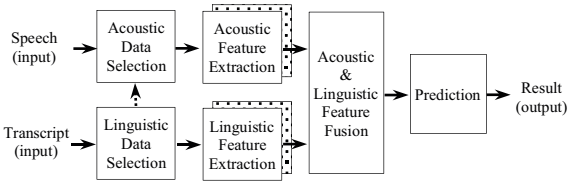

Figure 1: *Acoustic data selection involves chosen token words, whereas linguistic data selection uses all words. Feature selection was applied during the feature extraction stages.*

[1] Quality control listening per utterance was completed to check time-stamp accuracy (some transcriptions had time-stamp errors). File labels and their time-stamps are available per request via the authors.

4. Results

4.1. Token Words Versus Entire Utterances

The purpose of these experiments was to examine how smaller segments perform when compared with entire utterances, and determine which features or feature combinations contribute to better token word depression prediction. Formulaic filler word analysis nearly matched that of the entire utterances baseline when identically partitioned comparisons were made, as shown in Table 2. When evaluated against the entire utterances baseline, fillers gave lower overall MAE and RMSE.

Table 2: *Token word and entire utterances (baseline) depression prediction using eGeMAPS acoustic features and SVR.*

	Word/Phrase		Baseline (all utt.)	
Token Words	MAE	RMSE	MAE	RMSE
"*Hmm*"	3.85	5.07	5.22	6.90
"*Mhm*"	4.08	4.18	4.10	4.80
"*Mm*"	5.58	6.95	5.70	6.77
"*Uh*"	5.37	6.50	5.96	6.90
"*Umm*"	6.56	8.15	5.35	6.67
"*So*"	6.31	8.07	6.31	8.17
"*No*"	5.00	6.08	4.71	5.79
"*Yeah*"	7.17	8.87	6.10	7.10
"*Okay*"	5.73	6.74	5.55	6.54
"*You Know*"	8.67	9.92	3.07	4.31
All Average	**5.83**	**7.05**	**5.21**	**6.40**
Filler Average	**5.29**	**6.49**	**5.44**	**6.70**

Similar results were found for VoiceSauce features, but with lower error in general. These features may have performed better due to containing a greater number of total features than eGeMAPS. In addition, VoiceSauce applies more than one acoustic analysis method (i.e. Praat, Straight, Snack Sound Toolkit) for estimating frequency and energy formant-related features. VoiceSauce features, shown in Table 3, produced competitive results when compared with the entire utterances baseline.

Table 3: *Token word SVR depression prediction using VoiceSauce acoustic features*.

	Word/Phrase	
Token Words	MAE	RMSE
"*Hmm*"	2.91	4.13
"*Mhm*"	3.76	4.62
"*Mm*"	5.56	7.09
"*Uh*"	4.82	6.03
"*Umm*"	6.27	7.96
"*So*"	6.06	8.57
"*No*"	4.87	6.24
"*Yeah*"	7.38	10.91
"*Okay*"	5.08	6.68
"*You Know*"	7.52	8.96
All Average	**5.42**	**7.11**
Filler Average	**4.90**	**6.40**

Using these features, only one of the token words fillers "*umm*" achieved a higher MAE than that of the baseline. Note that "*hmm*" performed particularly well for depression prediction, generating a 2.91 MAE versus the 5.22 MAE in the baseline.

4.2. Linguistic Baseline System

Trends in the linguistic features were discovered for depressed speakers having higher range PHQ-8 scores (e.g. 15-23). For instance, depressed speakers tended to have a reduction in overall word syllable averages, reduced preposition usage, increased usage of pronouns, and overall simpler sentence structure based on average readability scores. While depressed versus healthy female speakers did not indicate a difference in average words per sentence, depressed males showed an overall reduction, especially for higher PHQ-8 scores. In experimenting with linguistic features the MAE and RMSE average using linguistic features was nearly equal to the entire utterances baseline acoustic functional features, 5.17 and 6.30, respectively.

4.3. Acoustic and Linguistic Features Combined

Experiments utilizing all acoustic features from token words along with linguistic features were completed in an attempt to attain the lowest the MAE and RMSE possible. The eGeMAPS, VoiceSauce, and linguistic features were concatenated as a single vector per utterance before prediction using SVR. In Table 4, experiments using a combination of acoustic token word and linguistic features produced the overall lowest MAE and RMSE average for fillers token words.

Table 4: *Token words SVR depression prediction using combined eGeMAPS, VoiceSauce, and linguistic features*.

	Combined	
Token Words	MAE	RMSE
"*Hmm*"	2.89	4.35
"*Mhm*"	3.45	4.63
"*Mm*"	5.90	7.13
"*Uh*"	5.00	6.32
"*Umm*"	6.18	7.82
"*So*"	5.05	7.09
"*No*"	5.06	6.31
"*Yeah*"	6.42	8.50
"*Okay*"	4.70	6.24
"*You Know*"	8.08	9.42
All Average	**5.27**	**6.78**
Filler Average	**4.75**	**6.22**

In the results presented to this point, only subsets of the training/test data could be used. To understand the depression prediction performance across the entire data, all token words were merged into training and test sets, which allowed for every speaker to be represented much like the baseline results found in [17]. Using the combined entire utterances baseline eGeMAPS features, prediction errors of 5.51 MAE and 6.83 RMSE were attained. When combined sets were then run using the filler words with eGeMAPS features, prediction errors of 6.07 MAE and 7.52 RMSE were achieved.

While the combined filler word results did not produce results as low as the entire utterances baseline, filler words are surprisingly accurate considering most comprise less than a second of speech. The combined filler word results may be an indication that some particular filler words and their acoustic-phonetic attributes are better for depression prediction than others.

4.4. *N*-Best Analysis

An *n*-best approach was experimented with, using the four lowest MAE/RMSE token words that, when combined, allowed score prediction for every test utterance; thus, creating a fair comparison with the entire utterances baseline. In Table 5, the best test MAE/RMSE performance was achieved using *n*-best eGeMAPS and linguistic features with feature reduction. The absolute improvement in MAE/RMSE over the entire utterances baseline was 0.95 and 1.24, respectively. The *4*-best token words ("hmm", "mhm", "no", "uh") were fillers and/or had nasal phonetic elements. Due to experimental time constraints, entire utterances baseline for VoiceSauce MAE/RMSE will be included in a later revised version. For token words, the VoiceSauce features attained similar results to eGeMAPS. However, they did not demonstrate further improvement with the addition of linguistic features and feature reduction.

Table 5: *Comparison of entire utterances baseline versus 4-best token words ("hmm", "mhm", "no", "uh") on all test speakers. Note * indicates feature selection applied.*

	eGeMAPS		VoiceSauce	
	MAE	RMSE	MAE	RMSE
All Utterances (similar to [17])	5.51	6.83	-	-
All Fillers	6.07	7.52	6.08	7.59
4-Best	4.72	5.76	4.71	5.71
4-Best + Linguistic	4.74	5.70	4.71	5.71
4-Best* + Linguistic*	4.56	5.59	4.71	5.71

5. Conclusion

This research demonstrates that thin slice speech data selection can be competitive for depression prediction when compared with using whole utterances. Moreover, results show that among the token words selected for study herein, fillers consistently provided the lowest depression score prediction error.

Filler words appear advantageous because they are naturally repeated in abundance. The general location of filler words is between phrase clauses; meaning they typically begin or end a phrase, making them potentially easier to identify with automatic speech recognition and/or keyword spotting applications. Future research, utilizing a larger set of filler words with equal counts and number of speakers could further help determine which specific fillers or phonetic content contains the most valuable speech information for depression prediction systems.

6. Acknowledgements

The work of Brian Stasak and Julien Epps was partly funded by ARC Discovery Project DP130101094. The work of Nicholas Cummins is supported by the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

7. References

[1] Ambady, N., & Rosenthal, R., "Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis", *Psych. Bulletin*, Vol. 111, No. 2, 256-274, 1992.

[2] Reynolds, D., & Rose, R., "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, 72-83, January 1995.

[3] Boakye, K., & Peskin, B., "Text-constrained speaker recognition on a text-independent task", *ODYSSEY '04*, The Speaker and Language Recognition Workshop, 2004.

[4] Shriberg, E., & Stolcke, A., "The case for automatic higher-level features in forensic speaker recognition", *INTERSPEECH*, Brisbane – Australia, 1509-1512, 2008.

[5] Ishihara, S., & Kinoshita, Y., "Filler words as a speaker classification feature", *SST 2010*, Melbourne – Australia, 34-37, 2010.

[6] Bridges, K., "Prosody and formulaic language in treatment-resistant depression: effects of deep brain stimulation", PhD Thesis, Steinhardt School of Culture, Education, and Human Development: NYU – USA, 2014.

[7] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T., "A review of depression and suicide risk detection and assessment using speech analysis", *Speech Communication*, Vol. 71, 10-49, 2015.

[8] Scherer, S., Lucas, G., Gratch, J., Rizzo, A., and Morency, J., "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews", *IEEE Trans. Affect. Comp.*, Vol. 7, 59-73, 2016.

[9] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., & Morency, L., "The distress analysis interview corpus of human and computer interviews", *LREC*, 3123-3128, 2014.

[10] Trevino, A., Quatieri, T., & Malyska, N., "Phonologically-based biomarkers for major depressive disorder", *EURASIP Journal on Advances in Signal Processing*, Vol. 42, 2011.

[11] Pope, B., Blass, T., Siegman, A., & Raher, J., "Anxiety and depression in speech", *Journal of Consulting and Clinical Psychology*, Vol. 35, 128-133, 1970.

[12] Sethu, V., Ambikairajah, E., & Epps, J., "Phonetic and speaker variations in automatic emotion classification", *INTERSPEECH*, ICSA, Brisbane – Australia, 617-620, 2008.

[13] Kroenke, K., Strine, T., Spitzer, R., Williams, J., Berry, J., & Mokdad, A., "The PHQ-8 as a measure of current depression in general population", *Journal of Affective Disorders*, Vol. 114, 163-173, 2009.

[14] Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing", *IEEE Transactions on Affective Computing*, in press, 2015.

[15] Shue, Y., Keating, P., Vicenik, C., & Yu, K., "VoiceSauce: a program for voice analysis", Proceedings of the *ICPhS* XVII, 1846-1849, 2011.

[16] Wu, D., Hanauer, D., Mei, Q., Clark, P., An, L., Lei, J., Proulx, J., Zeng-Treitler, Q., & Zheng, K., "Applying multiple methods to assess the readability of large corpus medical documents", *Student Health Technology Information*, 192, 647-651, 2013.

[17] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., & Pantic, M., "AVEC 2016 – depression, mood, and emotion recognition workshop and challenge", submitted June 2016.

[18] Cummins, N., Sethu, V., Epps, J., & Krajewski, J., "Relevance vector machine for depression prediction", in Proceedings of the *Annual Conference of the International Speech Communication Association*, *INTERSPEECH*, Dresdan – Germany, 110-114, 2015.