Analysis of Phonetic Markedness and Gestural Effort Measures for Acoustic Speech-Based Depression Classification

Brian Stasak
School of Elec. Eng. & Telecom.
University of New South Wales
Sydney, NSW, Australia
b.stasak@unsw.edu.au

Julien Epps
School of Elec. Eng. & Telecom.
University of New South Wales
Sydney, NSW, Australia
j.epps@unsw.edu.au

Aaron Lawson
Speech Tech. & Research Lab.
SRI International
Menlo Park, CA, USA
aaron@speech.sri.com

***Abstract*—While acoustic-based links between clinical depression and abnormal speech have been established, there is still however little knowledge regarding what kinds of phonological content is most impacted. Moreover, for automatic speech-based depression classification and depression assessment elicitation protocols, even less is understood as to what phonemes or phoneme transitions provide the best analysis. In this paper we analyze articulatory measures to gain further insight into how articulation is affected by depression. In our investigative experiments, by partitioning acoustic speech data based on lower to high densities of specific phonetic markedness and gestural effort, we demonstrate improvements in depressed/non-depressed classification accuracy and F1 scores.**

1. Introduction

Of the 300 million individuals globally afflicted by depression, it is estimated that fewer than half will receive treatment [1]. While dozens of clinical depression assessment methods include self-report and verbal interview, these approaches have an inherent degree of subjective bias. Moreover, in recent years, healthcare professionals have expressed doubts as to the overall effectiveness of current criteria guidelines for depressive disorder diagnosis [2]. Consequently, there is a crucial necessity for more objective diagnostic tools in helping identify and monitor individuals with depression. In the future, noninvasive affective sensing technology, such as automatic speech analysis tools, will assist physicians during clinical depression evaluations.

For at-risk patients, many clinical depression disorder assessments, such as the Mental State Examination (MSE) and Quick Inventory of Depressive Symptomatology (QIDS) require behavioral observations concerning a patient's verbal ability. Physicians often take into account a patient's rate of speech, volume, tonality, mood, and cognitive awareness during depression disorder evaluations [3]. Many depression disorder assessments contain sections for observational accounts related to patients' speech behaviors; a common symptom of clinical depression denoted as *psychomotor retardation* hinders an individual's cognition and fine motor control [3, 4]. Speech-based depression analysis literature has steadily reiterated that patients with depression audibly demonstrate abnormal speech patterns, including sluggish phonation rates, less articulatory precision, reduced prosody, decreased intensity, and atrophied speech intelligibility [4-7].

Recently, phono-articulatory complexity score-based systems have been investigated as a measure of articulatory kinematic effort for detection of pathological speech disease/disorder. Yet, in this literature, such as [8-12], only a small set of articulatory characteristics, speakers, and languages were examined. Furthermore, these investigations were primarily designed for speech pathology evaluation applications. It is conceivable that for depression, the loss of articulatory control and kinematics in individual's speech can also be measured using similar descriptive phonologically driven metrics.

A promising, practical and inexpensive method to evaluate articulatory-physiological speech parameters is analysis of phoneme occurrences/transitions based on spoken transcript analysis and acoustic-based features. Early key influential physiological articulation parameter investigations concentrated on articulatory descriptors, such as English vowel attributes [13], universal phonetic markedness [14, 15] and spoken articulatory gestures [16, 17]. These linguistically motivated investigations provided further insight on natural speech production, including static phoneme analysis and non-static articulatory transitions called articulatory gestures, which are essentially parameterized dynamic systems based on overlapping coordinated muscular movements. In [18, 19], a gestural score was proposed to help measure articulatory movements based on the linguistic order of articulatory change. Unlike a discrete sequence of phonemes, a gestural score has an advantage because it encodes a set of hidden states present in natural interconnected speech.

While analysis of articulatory measure aspects in spoken language-specific disorders/diseases is common, fewer studies exist that examine the impact of depression on individual's articulation, and perhaps more importantly, which phoneme or syllable combinations are most affected by depression. In [20], it was alleged that acoustic-based depression classification should place more emphasis on finding a correlation between depression severity and changes in articulatory acoustic phoneme parameters. [21] examined the correlation of individual phonemes and broad phoneme classes (i.e. vowels, consonants, nasals) to *psychomotor retardation* by individually evaluating average phoneme lengths and centroid

energy spread. However, in this study, several articulatory attributes were left unexplored (e.g. placement, roundness, tenseness). Moreover, dynamically, no investigative analysis of phonetic gestures including the degree of articulatory transformation from one phoneme to the next was considered. To date no investigation has examined the use of articulatory gestural effort in a speech-based acoustic feature depression classification framework.

In a previous work [22], it was discovered that the relative age of phoneme acquisition mastery could be used as a measure of articulation effort. This measure was used to partition spoken phrases into subsets from low to high articulation effort. The experimental results indicated that phrases with higher articulation effort generated higher depressed/non-depressed classification accuracy than lower ones. As a follow up extended analysis, we examine a different set of articulatory measures proximately grounded on phonetic markedness [14] and gestural efforts [18, 19]. Based on the aforementioned depression disorder speech behaviors [4-7], our hypotheses are:

- For depressed speakers, it is believed the symptom of psychomotor retardation, which reduces articulatory motor control, will exhibit itself to a greater extent in the production of some phoneme manners more than others.
- As changes in the demand for articulation gestural effort activation increases, depressed speaker characteristics will become acoustically more evident within these phrases.

The balance of this paper is organized as follows: Section 2 describes the experimental dataset. Section 3 contains the articulatory measures along with the acoustic-based system design. Lastly, Section 4 comprises of our experimental results and provides further discussion on how these measures can be applied during clinic depression assessment protocol.

2. Experimental Data

All experiments utilized a subset of the training and development from the Distress Analysis Interview Corpus (DAIC) [23]. The DAIC 'Wizard-of-Oz' subset utilized an animated virtual interviewer and its data collection protocol was explicitly designed for elicitation of behavioral data to assist diagnosis of psychological distress disorders (e.g. anxiety, depression, post-traumatic stress). This experimental subset was chosen because it has 82 female/male speakers, natural speech in a clinical room environment, a virtual human interviewer (providing neutral, unbiased emotion consistency in speech elicitation), a limited number of questions, phrase-level spoken transcriptions with time stamps, and a Patient Health Questionnaire (PHQ-8) score per speaker.

In addition to audio recordings in the DAIC, each speaker completed a PHQ-8 self-assessment [24]. The PHQ-8 consists of eight questions, which produces total scores between 0 and 24 (the higher the score the greater the depression severity). In our experimental research, two different classification speaker groups were analyzed based on a two restricted PHQ score ranges: non-depressed (0-4) and depressed (15-24) [similarly to 22, 25, 26]. Speakers within the mid-PHQ-8 range were purposely removed to better establish far-end class trends. According to the descriptors associated with the PHQ-8 scores [24], the non-depressed group can be considered to have "no current depression", whereas the depressed group has "current moderately severe to severe depression". Nearly 20% of speakers in this experimental subset were in the latter group (i.e. clinically diagnosed as depressed).

All phrases were sorted and partitioned into low, mid-low, mid-high, and high for the 17 individual phonetic markedness (M_k) and Hamming Distance (D^{HD}) measures. There were ~2,500 phrases per partition and all speakers had phrases represented within each partition.

3. Methods

3.1. Articulatory Effort Measures

DAIC transcripts were converted from standard text format to phoneme representations using the Carnegie Mellon University (CMU) phonetic dictionary [27]. This dictionary consists of phonetic spellings for over 130k American English words using 39 phoneme representations (see Table I - column one). Two different types of articulation measures were examined per phrase based on the Chomsky-Halle phonetic model [14]: (1) phonetic markedness; and (2) Hamming distance mean. In total there were 18 articulatory measures extracted per phrase. These articulatory measures were then used to partition phrases in low-to-high sorted order to determine which partitions produced the greatest discrimination between depressed and non-depressed speakers.

3.1.1 Phonetic Markedness

The phonetic markedness is a percentage-based distribution calculated by the following simple equation:

$$m_k = y_{i,k} \tag{1}$$

where m_k is the markedness of the k^{th} manner. Per utterance, the mean of m_k is represented by M_k. This percentage-based approach provided articulatory markedness distributions within each phrase, and allowed all different phrase lengths to be compared equally to each other. For example, according to Fig. 1, the name "*John*" /J-AO-N/ is represented by the three phonemes that have each of the following phonetic markedness distributions: voice (M_k = 1.0), back (M_k = 0.33), and anterior (M_k = 0.66).

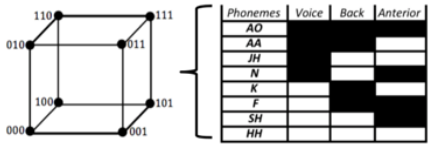

Figure. 1. Illustration of the Hamming distance for a 3-bit string. The Hamming distance between 000 and one of 100, 010, or 001 is only 1, whereas there is a distance of 3 between 000 and 111. This was applied to calculate the articulatory distance between consecutive phoneme pairs, by calculating the distances between the relevant pair of rows found in Table 1.

3.1.2 Hamming Distance Mean

Considering Table 1, the articulatory gestural effort due to the *transition* between a pair of consecutive phonemes can be reflected by the difference in the Chomsky-Halle phonetic markedness corresponding to each parameter. Since the Chomsky-Halle

parameters are binary, this is a bitwise distance. The Hamming distance [28] is a bitwise metric that measures the minimum number of substitutions required to transform from one binary string to another binary string of equal length. It can be represented by the following equation:

$$d^{HD}(i,j) = \sum_{k=0}^{n-1} |y_{i,k} - y_{j,k}| \quad (2)$$

where the value difference between y_i and y_j is determined by the index k based on the total number of variables n. The Hamming distance gives the number of mismatches between variables paired by k, and in this context, is related to how many changes in manner occur during a particular sequence of phonemes. This is treated herein as a proxy for the articulation effort of phoneme transitions. In Fig. 1, the higher the Hamming distance, the greater the difference between two binary string representations. The illustrated example previously shown in Fig. 1, indicates the phonetic markedness m_k within phoneme /HH/ is least like /AO/, hence the larger gestural effort (d^{HD} = 1.0) than similar phonemes /SH, JH, K/ (d^{HD} = 0.66).

TABLE I. ARTICULATION EFFORT MATRIX BASED ON CHOMSKY-HALLE'S 17 PHONETIC MARKEDNESS [14]. THE COLUMNS (k) ARE THE MANNER AND THE ROWS (i) ARE ENGLISH PHONEMES; WHITE INDICATES *INACTIVE* ($y_{i,k}$ = *0*), WHEREAS BLACK INDICATES *ACTIVE* ($y_{i,k}$ = *1*). THE PARENTHESES INDICATE TOTAL NUMBER OF PHONEMES CONTAINED WITHIN.

		Vocalic (15)	Consonantal (25)	High (13)	Back (9)	Low (4)	Anterior (21)	Coronal (14)	Round (7)	Tense (7)	Voice (30)	Continuant (27)	Nasal (3)	Strident (8)	Sonorant (22)	Interrupted (8)	Distributed (6)	Lateral (1)
odd	AA	■			■	■					■	■			■			
at	AE	■				■	■				■	■			■			
hut	AH	■		■							■	■			■			
ought	AO	■			■	■	■			■	■	■			■			
cow	AW	■					■			■	■	■			■			
hide	AY	■					■				■	■			■			
Ed	EH	■					■				■	■			■			
ate	EY	■					■				■	■			■			
it	IH	■		■							■	■			■			
eat	IY	■		■			■			■	■	■			■			
oat	OW	■			■					■	■	■			■			
toy	OY	■								■	■	■			■			
hood	UH	■			■						■	■			■			
two	UW	■		■	■					■	■	■			■			
hurt	ER	■	■				■	■			■							
be	B		■								■					■		
dee	D		■				■	■			■					■		
pee	P		■				■									■		
tea	T		■				■	■								■		
vee	V		■				■				■	■		■			■	
zee	Z		■				■	■			■			■			■	
seizure	ZH		■	■				■	■		■			■				
sea	S		■				■	■				■		■				
she	SH		■	■				■				■		■			■	
me	M		■				■				■		■		■			
knee	N		■				■	■			■		■		■			
ping	NG		■	■							■				■			
thee	DH		■				■	■			■	■						
cheese	CH		■	■				■						■		■	■	
fee	F		■				■					■						
go	G		■	■							■					■		
he	HH		■			■						■			■		■	
jeep	JH		■	■				■			■	■		■		■		
key	K		■	■														
Lee	L		■				■	■			■	■			■			■
read	R		■				■	■			■	■			■			
theta	TH		■				■	■				■					■	
we	W		■	■					■		■	■			■		■	
yield	Y		■	■							■	■			■			

During experimental work, in an identical fashion to Fig. 1, each phoneme was represented as a 17-bit string (see Table I). For every phrase a Hamming distance mean was calculated by calculating d^{HD} between all consecutive phoneme pairs across all 17 phonetic markedness articulation manners. After a Hamming distance measure according to M_k for each k^{th} manner, D^{HD} was calculated as the mean of all d^{HD} for the phrase.

3.2. Acoustic System Configuration

For experiments herein, the open-source COVAREP speech toolkit [29] was used to extract 74 acoustic features (i.e. glottal flow, MFCC, pitch) by aggregating 20-ms frame-level features across individual utterances. Each feature had its mean, standard deviation, kurtosis, and skewness calculated, which generated a total feature vector dimensionality of 296. For a baseline, each speaker's acoustic features were averaged based on all phrases without partitioning. During the partitioning experiments, the acoustic features corresponding to only those speaker phrases belonging to the identical partition were averaged. The COVAREP feature set was chosen because it has been used previously for emotion and speech-based depression research on this dataset [30, 31].

Similarly to [32], depression classification was conducted using decision trees, which performed well in preliminary experiments. All experiments used the simple decision tree classification from the MATLAB toolkit using a few leaves with coarse decision-making and a maximum of 4 splits. Experiments utilized 2-class classification (depressed/non-depressed) with 5-fold cross validation using a 20/80 training/test split. Classification performance was determined using overall accuracy, individual class F1 scores, and F1 average scores [similarly to 30]. The F1 score is a common metric that determines precision and recall based on true/false positives and true/false negatives, and is a helpful evaluation criterion for unbalanced classification problems. The F1 score is computed as follows (a large F1 score implies better discrimination):

$$F1 = \frac{precision \cdot recall}{precision + recall} \quad (3)$$

4. Results & Discussion

The baseline acoustic feature experiment resulted in an average accuracy of 69.5% with F1 scores of 0.07(0.82), i.e. depressed and non-depressed, respectively. In comparison to the baseline, results in Fig. 2(a) show depressed/non-depressed classification accuracy performance for each static phonetic markedness parameters for low (light gray), mid-low, mid-high, and high (dark gray) phrase partitions. In regards to our first hypothesis, for several phonetic markedness parameters, the results indicate that depressed/non-depressed accuracy is better for higher than lower density phonetic markedness partitions. This is particularly the case for consonantal, round, and voice phonetic markedness, wherein classification accuracy improved as these parameters had greater prominence in phrases. Surprisingly, the opposite effect was recorded with strident and lateral phonetic markedness parameters – where phrases with fewer of these parameters produced higher classification accuracy.

Fig. 2(b) shows the F1 scores for depressed for individual phonetic markedness parameters. The F1 depressed classification

results in Fig. 2(b) display a great deal of variance in scores across nearly every phonetic markedness parameter, whereas for Fig. 2(c) the F1 non-depressed classification score ranges are significantly narrower. The depressed F1 scores are considerably lower than the non-depressed F1 scores. One reason for this difference in results could be the 1:5 ratio of depressed to non-depressed speakers, which means significantly less training data available for the depressed than the non-depressed class. Preliminary experimental results (not shown here) using balanced speaker classes showed classification improvements for depressed F1 scores, albeit with minor costs in classification accuracy. However, from a clinically realistic sample standpoint, moderately severe to severe depressed speakers typically make up a smaller percentage of the population [33].

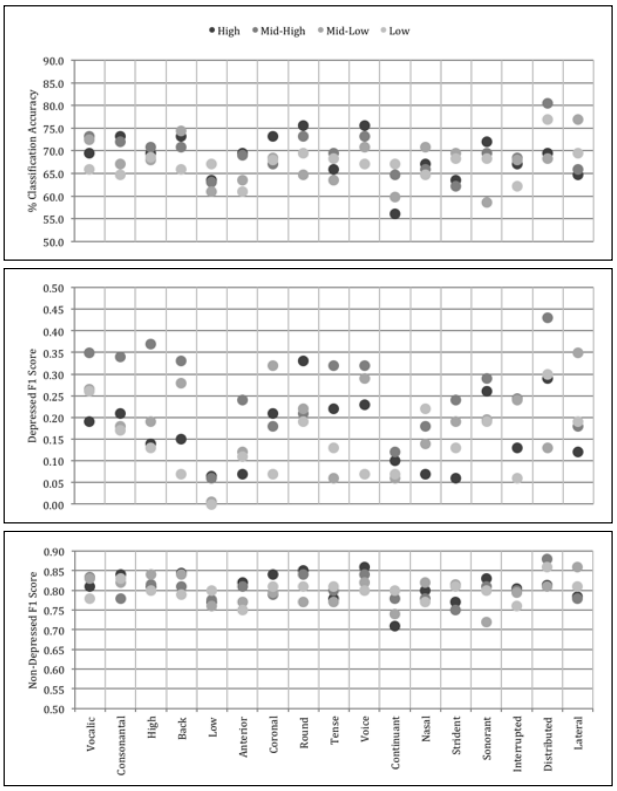

Figure. 2. (a) Classification accuracy, (b) depressed F1 score, and (c) non-depressed F1 score results for acoustic features based on sorted mean phrase activation (M_k) of all 17 phonetic markedness parameter partitions.

Utilizing the Hamming distance (D^{HD}) to partition speakers' acoustic features per phrase, Fig. 3 results show absolute accuracy gains of around 5% when using mid-high to high partitions rather than low to mid-low partitions. While the absolute accuracy gain is only a small improvement for higher Hamming distance partitions over the all phrase baseline (~2%). However, there were considerable depressed F1 score absolute accuracy gains for the higher partitions (to 0.21). As suggested by our second hypothesis, the depressed F1 score improvement indicates that a greater number of depressed speakers are correctly classified using acoustic features from phrases that have a larger Hamming distance mean. These improvements found with the mid-high and high Hamming distance partitions suggest that elicitation of potentially depressed patients' speech should contain the widest phoneme-to-phoneme activation range of articulatory gestural transitions.

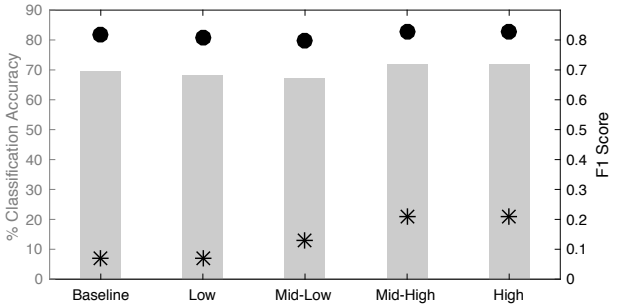

Figure. 3. Classification accucracy, depressed (star) and non-depressed (dot) F1 score performance results for the baseline (all phrases) versus phrase phonetic markedness D^{HD} Hamming distance partitions.

Differences in the number of phonemes per phonetic markedness parameters suggest that if the 17 different articulatory marked parameters were weighted, the results might achieve better classification performance, especially in regards to the additional Hamming distance metric. But, the issue of determining suitable weights per articulatory trait is a problem found in many of the score-based systems mentioned earlier in Section I [8-12]. Moreover, there is surprisingly a scarcity of literature related to the kinematic demand and processes of speech [34-36]. While it is presumed that greater muscular involvement equals greater articulatory effort, there is however, no literature for instance that provides a quantitative guide on whether the rounding of the lips requires less, equal, or more effort than voicing a phoneme. To gain greater insight into speech affected by depression along with other common disease/disorders, additional research in the area of measurable phono-articulatory kinematics is needed.

Although the experiments herein rely heavily on human transcribed speech, current Automatic Speech Recognition (ASR) software is achieving near-human word error rates. It has also been suggested in [37, 38] that hidden gestures or articulators could be automatically identified using speech recognition algorithms. Hence, rather than a standard phoneme to readable written text-transcript output, the output contains non-text gestural or articulator information. As for collecting a diverse set of phonemes and transitions, one obvious articulatory limitation is the naturally occurring frequency of phonemes in spoken English.

5. Conclusions

Our experimental approach using Chomsky-Halle phonetic markedness and the Hamming distance mean shows initial promise as an expanded area research for speech-based depression analysis. Most interesting is the implication of greater articulatory knowledge on depressive disorder elicitation protocol design and speech data selection. By better understanding what speech sounds and

combinations are altered by depression, more effective elicitation protocols for speech can be developed to futher ameliorate automatic acoustic analysis performance and reliability. To help counter naturally occurring phoneme distributions found across all languages, specifically designed read excerpts could be used for depression assessments that contain the most salient and discriminant phonetic markedness and/or transitions. This would thereby guarantee a higher degree of articulatory demand, resulting in improved depression classification accuracy.

6. Future Work

Likewise to the phonetic markedness parameters investigated herein for depression, other levels of prominence should be analyzed, such as lexical stress, vowel reduction, and pitch contours. It is suggested that these phonetic markedness measures might also work well for automatic speech-based classification severity of Parkinson's disease, especially noting the abnormal articulatory productions and manner characteristics found in [39].

Acknowledgments

The work of Brian Stasak and Julien Epps was partly funded by ARC Discovery Project DP130101094 and research supported by Data61, Sydney - Australia.

References

[1] World Health Organization (WHO), *Factsheet* Feb. 2017: http://www.who.int/mediacentre/factsheets/fs369/en/

[2] J. Doward, "Medicine's big new battleground: does mental illness really exists?", *The Guardian Newspaper*, May 12th 2013.

[3] D. Bennabi, P. Vandel, C. Papaxanthis, T. Pozzo, & E. Haffen, 'Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications', *BioMed Research International*, Vol. 2013, 2013.

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

[5] K. Scherer, & B. Zei, 'Vocal indicators of affective disorders', *Psychother Psychosom*, Vol. 49, pp. 179-186, 1988.

[6] A.J. Flint, S.E. Black, I. Campbell-Talor, G.F. Gailey, & C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression', *J. of Psych. Research*, Vol. 27(3), pp. 309-319, 1993

[7] J. Darby, N. Simmons, & P. Berger, 'Speech and voice parameters of depression: a pilot study', *J. Commun. Disord*., Vol. 17, pp. 75-85, 1984.

[8] C. Stoel-Gammon, "The word complexity measure: description and application to developmental phonology and disorders", *Clinical Linguistics & Phonetics*, Vol. 24(4-5), pp. 271-282, April-May 2010.

[9] K. Jakielski, R. Maytasse, & E. Doyle, "Acquisition of phoentic complexity in children 12-36 months of age", poster presented at the convention of the American Speech-Language-Hearing Association, Miami, FL – USA, 2006.

[10] A.W. Maccio, "Clinical problem solving: assessment of phonological disorders", *American J. of Speech-Language Pathology*, Vol. 11, pp. 221-229, 2002.

[11] L. Shriberg, D. Austin, B. Lewis, J. McSweeny, & D. Wilson, "The percentage of consonants correct (PCC) metric: extensions and reliability data", *Journal of Speech, Language, and Hearing Research*, Vol. 40, pp. 708-722, 1997.

[12] L.I. Shuster, & C. Cottrill, "Ease of articulation: a replication", *Biotechniques*, Vol. 22(5), pp. 952-957, May 1997.

[13] P. Ladefoged, A Course in Phonetics, Harcourt Brace Jovanovich, 3rd Ed., 1993.

[14] N. Chomsky, & M. Halle, The Sound Pattern of English, Harper & Row, New York, NY - USA, 1968.

[15] A. Henning, "Markedness – the first 150 years", Markedness in Synchrony and Diachrony, Olga M. Tomic (ed.), Mouton de Gruyter, Berlin – Germany, pp. 11-46, 1989.

[16] C.P. Browman & L. Goldstein, "Towards an articulatory phonology", *Phonology Yearbook*, Vol. 3, pp. 219–252, 1986.

[17] C.P. Browman & L. Goldstein, "Articulatory phonology: An overview", *Phonetica*, Vol. 49, pp. 155–180, 1992.

[18] C.P. Browman & L. Goldstein, "Articulatory gestures as phonological units", *Phonology*, Vol. 6, pp. 201–250, 1989.

[19] C.P. Browman & L. Goldstein, "Dynamics and articulatory phonology", In Port, R. and v. Gelder, T. (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition, pp. 175–193. MIT Press, 1995.

[20] K. Gábor & V. Klára, "Physiological and cognitive status monitoring on the base of acoustic-phonetic speech parameters", in L. Besacier, A.H. Dediu, & C. Martin-Vide (eds), Statisitical Lang. and Speech Processing, 2014.

[21] A. Trevino, T. Quatieri, & N. Malyska, "Phonologically-based biomarkers for major depressive disorder", *EURASIP Journal on Advances in Signal Processing*, pp. 1-18, 2011.

[22] B. Stasak, J. Epps, & R. Goecke, "Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect", *InterSpeech '17*, Stockholm – Sweden, *in press* August 2017

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

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

[25] C. Soloman, M.F. Valstar, R.K. Morriss, & J. Crowe, "Objective methods for reliable detection of concealed depression", *Frontiers In*, Vol. 2(5), pp. 1-16, April 2015.

[26] Z. Liu, B. Hu, F. Liu, H. Kang, X. Li, L. Yan, & T. Wang, "Evaluation of depression severity in speech", Inter. Conf. on Brain and Health Informatics, Springer International Publishing, pp. 312-321, 2016.

[27] CMU (1993). The Carnegie Mellon Pronouncing Dictionary v0.1. Carnegie Mellon University: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[28] R.W. Hamming, "Error detecting and error correcting codes", Bell System Technical Journal Vol. 26(2), pp. 147-160, 1950.

[29] G. Degottex, J. Kane, T. Drugman, T. Raitio, & S. Scherer, "COVAREP – a collaborative voice analysis repositiory for speech technologies", in *Acoustics, Speech and Signal Processing* (ICASSP), pp. 960-964, 2014.

[30] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M.T. Torres, S. Scherer, G. Stratou, R. Cowie, & M. Pantic, "AVEC 2016 – depression, mood, and emotion recognition workshop and challenge", Proc. of the 6th Inter. Workshop on Audio/Visual Emotion Challenge, Amsterdam – The Netherlands, pp. 3-10, Nov. 2016.

[31] N. Cummins, B. Vlasenko, H. Sagha, & B. Schuller, "Enhancing speech-based depression detection through gender deprendent vowel-level formant features", in Teije, A., Popow, C., Holmes, J., & Sacchi, L., (eds) Artificial Intelligence in Medicine, AIME 2017, Lecture Notes in Computer Science, Vol. 10259, Springer, Cham, 2017.

[32] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, & M. Graciarena, 'The SRI AVEC-2014 evaluation system', *AVEC '14*, Orlando, FL – USA, pp. 93-100, Nov. 2014.

[33] World Health Organization (WHO), "The World Health Report 2001: Mental Health, New Understanding, New Hope", Geneva, 2001.

[34] J. Locke, "Cost and complexity: selection for speech and language", *Journal of Theoretical Biology*, Vol. 251, pp. 640-652, 2008.

[35] F. Pellegrino, E. Marsico, I. Chitoran, & E. Coupe (eds), Approaches to Phonological Compexity, Phonology and Phonetics Series, Vol. 16, Walter de Gruyter: Berlin, Dec. 2009.

[36] J.R. Westbury & J. Dembowski, "Articulatory kinematics of normal diadochokinetic performance", *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, Vol. 27, pp. 13-36, 1993.

[37] K. Livescu, & J. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities", In *ICSLP- 04*, Jeju - South Korea, 2004.

[38] K. Livescu, "Feature-based pronuncaition modeling for automatic speech recognition", Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[39] J.A. Logemann & H.B. Fisher, "Vocal tract control in Parkinson's disease: phonetic feature analysis of misarticulations", *Journal of Speech and Hearing Disorders*, Vol. 46, pp. 348-352, 1981.