INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Canonical Correlation Analysis and Prediction of Perceived Rhythmic Prominences and Pitch Tones in Speech*

Elizabeth Godoy, James R. Williamson, Thomas F. Quatieri

MIT Lincoln Laboratory, 244 Wood Street, Lexington MA, 02420 USA
elizabeth.godoy@ll.mit.edu, jrw@ll.mit.edu, quatieri@ll.mit.edu

Abstract

Speech prosody encodes information about language and communicative intent as well as speaker identity and state. Consequently, a host of speech technologies could benefit from increased understanding of prosodic phenomena and corresponding acoustics. A recently developed comprehensive prosodic transcription system called RaP (Rhythm-and-Pitch) annotates both perceived rhythmic prominences and pitch tones in speech. Using RaP-annotated speech corpora, the present work analyzes relationships between perceived prosodic events and acoustic features including syllable duration and novel measures of intensity and fundamental frequency. Canonical Correlation Analysis (CCA) reveals two dominant prosodic dimensions relating the acoustic features and RaP annotations. The first captures perceived prosodic emphasis of syllables indicated by strong metrical beats and significant pitch variability (i.e. presence of either high or low pitch tones). Acoustically, this dimension is described most by syllable duration followed by the mean intensity and fundamental frequency measures. The second CCA dimension then primarily discriminates pitch tone level (high versus low), indicated mainly by the mean fundamental frequency measure. Finally, within a leave-one-out cross-validation framework, RaP prosodic events are well-predicted from acoustic features (AUC between 0.78 and 0.84). Future work will exploit automated RaP labelling in contexts ranging from language learning to neurological disorder recognition.

Index Terms: speech prosody, rhythmic prominence, prosodic acoustics, Canonical Correlation Analysis

1. Introduction

Prosody describes the multi-tiered organization of sounds in spoken language that conveys meaning through emphasis and intonation [1]. This meaning can focus listener attention to specific speech content, through word or phrasal accentuation, in addition to reflecting speaker attitude, identity, and emotional state. Though rich in information, prosody is complex and represents a challenging area of investigation. Prosodic research spans an inter-disciplinary mix of linguists, cognitive scientists and engineers. However, speech technology applications, including recognition (automatic speech, language, speaker, neurocognitive state) and synthesis contexts often treat prosody with limited capacity due to scarcity of data with expertly-labeled percepts [2].

* This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

Leveraging recent linguistic advances in transcribing perceived rhythmic prominences, this work provides a comprehensive statistical framework for acoustic analysis and prediction of speech prosodic events.

In the past decade, linguistic researchers have successfully developed a transcription system for speech prosody that captures both perceived rhythmic prominences and pitch tones. Analogous to music, perceived rhythmic prominences reflect patterning of metrical beats in speech. Though these metrical prominences are often related to pitch tones, or tonal patterns, they are not limited to pitch accents. This distinction, namely that tone is not a requisite for perceived rhythmic prominence, motivated creation of the Rhythm and Pitch (RaP) transcription system [3] to improve upon Tones and Break Indicators (ToBI) annotation [4]. In inter-transcriber reliability studies, the RaP system proved a viable alternative to ToBI, while additionally annotating perceived metrical beats [5]. The following work uses RaP-transcribed English speech corpora to characterize and predict acoustics of the perceived prosodic events.

Across disciplines, the acoustics of prosody are typically considered to be composed of sound duration, fundamental frequency (f0) and intensity [1, 3, 4, 6, 7], though the definition of the sound unit is variable. While pitch and energy measures are calculated on a segmental (i.e. speech frame) level, prosodics are supra-segmental. The length of sounds spans a hierarchy, beginning with the shortest units: frame, phone, syllable, word, phrase, utterance, conversation. Syllables, rather than hierarchically adjacent phones and words, provide a perceptually salient framework for prosodic analyses. Syllables have been shown to be the most perceptually relevant unit in characterizing different speaking rates [8-10] and provide a common rhythmic unit describing linguistic typologies and metrical patterning [11]. While the present work uses RaP data with hand-annotated syllables, pseudo-syllable detection (focused on targeting vowels with grouping of surrounding consonants) offers an unsupervised method for estimating syllabic units [12].

In the present work, acoustic analyses of the syllable percepts begin with extraction of novel features for fundamental frequency (f0) and intensity (energy) respectively based on locally mean-normalized f0 contours and iteratively-estimated time domain amplitude envelopes. Statistically, a notable contribution of this work is to then apply canonical correlation analysis (CCA) in a new domain [13-15], specifically to relate the prosodic labels (outcomes) and acoustic measures (features). CCA linearly projects multiple variables, across feature and outcome spaces in this prosodic context, to a coordinate system such that their mutual correlation is maximized. Particularly well-suited to capture dependencies between multi-variate data, CCA is used specifically here to

http://dx.doi.org/10.21437/Interspeech.2017-1585

reveal the relative strctures of, and relationships between, the annotated rhythmic prominences and pitch tones with respect to duration, fundamental frequency and intensity features. Finally, using a CCA-based prediction model with leave-one-out cross validation, labeled prosodic events are predicted well (AUC around 0.8) from the syllable acoustics.

The remainder of this paper is structured as follows. Section 2 describes the RaP transcription and annotated corpora. Section 3 then details the acoustic feature extraction. Section 4 subsequently describes the canonical correlation analysis and prediction results. Conclusions are discussed in Section 5.

2. RaP Transcription Data

2.1. Transcription Labels

Rhythm and Pitch (RaP) transcription of prosody is broken down into two principal tiers. First for the Rhythm tier, prominence annotations are made on the syllable level based on perception of metrical beats. Second for the Pitch tier, tones are labelled sequentially using perception assisted by signal (waveform, pitch contour) displays. Though typically occurring one-per-syllable, RaP privileges annotator perception and does not explicitly constrain tone labeling in this way. More details on the labeling methodology and inter-transcriber reliability can be found in the work of Dilley et al [3,5]. The focus of this work is on the primary rhythmic prominence and pitch tone category labels outlined in Table 1.

Table 1: *Rhythm and Pitch (RaP) Transcription.*

Rhythm	
X	Strong metrical prominence
x	Weak or moderate metrical prominence
nXx	No prominence label
Pitch	
H	High tone
L	Low tone
E	Equal tone
nT	No tone label

2.2. Transcribed Speech Corpora

Two RaP-annotated English speech corpora, provided to the authors by Laura Dilley and Maura Breen, are analyzed in this work [5]. The first is 13-minutes of annotated telephone conversations (17 recordings in total) taken from the CALLHOME American English Speech corpus [16]. The second is 14-minutes of annotated radio broadcast news (28 recordings in total) taken from the BURSC-Boston University Radio Speech Corpus [17]. Further details on the corpora and annotation can be found in [16, 17] and [5], respectively. With limited annotated data available, analysis results in this work consider the union of these corpora, 45 recordings in total.

The rhythmic prominence and pitch tone label distributions for the RaP-annotated data are shown in Figure 1. Plotted on the left are conditional distributions of the pitch tone labels given the rhythmic prominence. Complementary distributions (prominence given tone label) are plotted on the right. Examining Figure 1, a clear trend is apparent. First, syllables with no beats and no tones co-occur frequently. Second, on the other end of the prosodic spectrum, syllables with major beats and primarily high pitch tones co-occur frequently. These distributions indicate the following syllable percept: the stronger the determinant of metrical beats, the more likely pitch is highly variable (high or low pitch tones are perceived). Conversely, syllables that have little influence on rhythmic beats typically contain little pitch variability (i.e. no or equal tones).

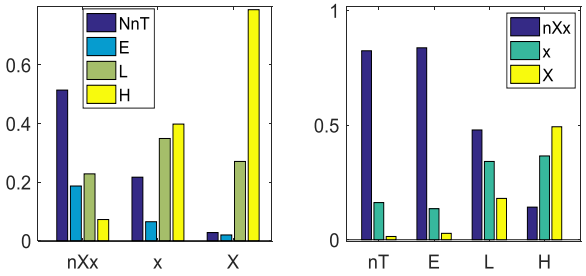

Figure 1: *RaP label conditional distributions for the spontaneous speech corpora (left: Pitch given Rhythm label, right: Rhythm given Pitch label).*

3. Acoustic Feature Extraction & Analyses

Metrics for analyses and quantification of speech rhythm often focus on duration [18]. In this work, syllable duration is calculated from the annotated start and stops times provided in the data from Dilley et al [5]. In addition to duration, acoustic measures of fundamental frequency (f0) and intensity are also extracted and calculated on the syllable level. In total, five syllable features are considered: duration plus mean and standard deviation (std) of both f0 and intensity.

3.1. Fundamental Frequency Measure

The fundamental frequency measure is based on a processed Praat contour with outlier mitigation and local-mean normalization. First, the Praat pitch and voicing provide the raw f0 values [19]. Next, to mitigate the effects of pitch doubling and halving, the top and bottom 5% of f0 values for each file are zeroed and subsequently replaced by the nearest neighbour interpolated f0 value [20]. Finally, because the pitch tone labels are locally relative (chosen by comparing perceived pitch to that of the previously labelled syllable), local detrending is done by subtracting a smoothed f0 contour calculated via convolution with a 750ms Hanning window. Unvoiced frames are replaced by nearest-neighbor interpolated values in the smoothing. Figure 2 plots an example of the resulting f0 measure for a major beat syllable and its precedent. For display purposes, the signal, f0, and intensity measures have each been normalized so the recording maximum absolute value is 1. In Figure 2, the f0 measure is clearly high for the major beat syllable with an "H" marked tone and low for the preceding syllable with an "L" marked tone. Finally, only voiced frames are used to calculate the syllable-based features, mean and std f0.

3.2. Intensity Measure

The intensity measure is based on an iterative time-domain signal envelope estimation that provides a smooth contour of amplitude peaks [21]. This estimation technique is inspired by the "True" spectral envelope described in [22]. First, the signal absolute value is convolved with a 10ms moving average filter. Peaks (local maxima) are detected and if none occur within 50ms of each other, or a max number (150) of iterations has been reached [21], the smoothed signal is the final output. Otherwise, the next iteration starts with the output of a max operation applied between the low pass

filtered and original absolute signal data. An example of the estimated time-domain envelope is shown in Figure 2. The final intensity measures are the mean and std of these envelope values per syllable.

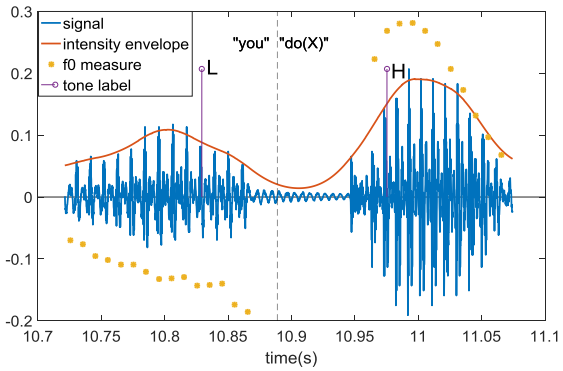

Figure 2: *RaP-transcribed syllables "you do" where "do" is a major rhythmic beat-X. The waveform is shown with intensity (red -) and fundamental frequency (orange *) normalized measures.*

4. Analysis and Prediction Results

The aim of the following analyses is to uncover underlying relationships between the five acoustic features described in Section 3 and the discrete prosodic RaP label outcomes (listed in Table 1) for the speech corpora. Canonical Correlation Analysis (CCA) is well-suited for this prosodic characterization as it has been shown in other domains to perform well in discovering associations between multivariate data sets [14-15]. The following describes application of CCA for acoustically-based prosodic discovery and prediction.

4.1. Canonical Correlation Analyses (CCA)

Given two multivariate vectors, X and Y, CCA finds linear combinations of the X_i and Y_j, projecting them to one-dimensional variables that maximize mutual correlation [13]. This is done iteratively to maximally account for residual correlation in each successive CCA projection dimension. CCA thereby reveals multivariate relationships among two sets of variables occupying different coordinate spaces.

In the following analysis, CCA is applied to two multivariate vectors: 1) the five dimensional acoustic features, and 2) the binary-valued RaP label outcomes. Before applying CCA, variables are first normalized using z-scoring so that relative contributions can be interpreted on the same magnitude scale.

Table 2 provides the Pearson correlations R across the CCA dimensions between the projected acoustic features and projected RaP outcome vectors. Clearly, the first two CCA components dominate the analyses, with CCA-3 (0.14) yielding an R less than half of that of CCA-2 (0.37). These R values suggest that there are two principal axes on which the outcome and feature data are related.

For the two dominating components, CCA-1 and CCA-2, the corresponding projection weights for the outcomes and features are provided in Table 3. First for CCA-1, the rhythm tier follows a clear hierarchy: nXx-no prominence has a weight of 0, x-weak metrical beat a weight of 0.36 and X-strong metrical beat weight a weight of 0.61. While the RaP transcription inherently specifies this hierarchy, CCA has statistically uncovered the relationship and quantified its influence on the dominant prosodic dimension. For the pitch tone tier in CCA-1, a clear hierarchy also emerges along pitch variability: nT-no tone and E-equal are comparable while L-low and H-high tones are doubly weighted, with H dominating slightly over L. This weighting indicates that pitch variability (presence of H or L tones) is more significant in this principle dimension than tone level. Acoustically, feature weights in the CCA-1 dimension are dominated by syllable duration, followed moderately by mean intensity and fundamental frequency (f0). Next, the secondary CCA-2 dimension is largely captured by the pitch tier in the outcome space, specifically L, with H having almost zero weight. This projection is thus describing the pitch tone level (low-equal-high), with H and L on opposite ends of the hierarchy. Acoustically for CCA-2, the mean f0 weight dominates, followed by duration and std of intensity. Together, Tables 2 and 3 indicate that the major axis of correlation encodes the presence/absence of prosodically emphasized syllables (i.e. those having strong metrical beats and high pitch variability) captured primarily by duration followed by mean f0 and intensity. The subsequent axis then captures a secondary pitch tone level dimension indicated primarily by mean f0.

Table 2: *Outcome-Feature Correlations (R) across CCA dimensions.*

CCA	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
R	0.56	0.37	0.14	0.08	0.02

Table 3: *CCA-1, CCA-2 outcome and feature weights.*

Outcomes	CCA-1 wts	CCA-2 wts
X	0.61	-0.14
x	0.36	0.12
nXx	0	0
H	0.75	0.03
L	0.67	1.13
E	0.32	0.35
nT	0.33	0.31
Features	CCA-1 wts	CCA-2 wts
Duration	0.91	0.52
Intensity mean	0.22	0.15
Intensity std	-0.02	-0.32
F0 mean	0.27	-0.91
F0 std	0.11	-0.10

4.2. Prediction Results

The following cross-validation analysis evaluates how well the CCA transform generalizes to held-out test data. Results are given in terms of the correlations of CCA projections from the acoustic features and RaP outcome labels. Moreover, since the RaP labels consist of two distinct tiers (Rhythm and Pitch), predictive generalization is evaluated for each tier individually and for both tiers combined. Additionally, leave-one-out cross-validation was used, testing on each recording after training on the other 44 recordings. The summary statistic used to assess predictive generalization is the mean Pearson R value of the acoustic-based projection with the RaP label-based projection.

Table 4 summarizes the predictive generalization results. The RaP tier is indicated by the first column. The second and third columns provide the global R values on all of the corpora data for CCA-1 and CCA-2, broken down by Tier. The fourth and fifth columns show the average R values obtained on the test data for CCA-1 and CCA-2 in the cross-validation framework, which are comparable to the Global results. Table 4 indicates that the Rhythm tier impacts the CCA-1 component (i.e. the R values for CCA-2 are relatively small). On the other hand, for the Pitch tier, both the CCA-1 and CCA-2 components are strong, as indicated by the high R values. In other words, the Pitch tier plays a significant role in both CCA-1 and CCA-2 dimensions, while the Rhythm tier primarily impacts the dominant CCA-1 dimension. In combination, the Rhythm and Pitch together always increase the R values, indicating that the tiers are providing complementary prosodic information. This observation supports the RaP premise that prominences and tones, while related, are not causally dependent.

One note in analyses concerns results examined for the two, CALLHOME and BURSC, corpora. Correlations and weights for the corpora essentially matched their union results, with one notable exception: CCA-1 and CCA-2 R values for the Pitch tier were especially high (0.45-Global) for BURSC compared to CALLHOME (0.24-Global), suggesting the radio broadcast style may exaggerate pitch, at least more than in casual conversation, to presumably emphasize news content.

Table 4: *CCA Global & Cross Validation (CV) Prediction Results by RaP tier.*

	Global R values		CV Mean R values	
RaP Tier	CCA-1	CCA-2	CCA-1	CCA-2
Rhythm	0.52	0.10	0.56	0.09
Pitch	0.50	0.36	0.51	0.43
Both	**0.56**	**0.37**	**0.59**	**0.44**

Additionally, in order to visualize the relationship between the CCA dimensions and RaP outcomes, Figure 3 plots one-half standard deviation contours of the projected outcome vectors in both CCA dimensions, broken down by rhythm and pitch label category. Strength of rhythmic prominences (line style) is shown to significantly influence the predicted outcomes, especially along the CCA-1 dimension. In other words, the dotted lines (no prominence) are clustered to the left of Figure 4, while the solid thick lines (major beats) are localized to the right, with dashed (moderate beats) lying in the middle. Considering the tones, both CCA-1 and CCA-2 dimensions effectively cluster different tone levels: this is most clearly seen with significant pitch variability (i.e. H&L, H, L) clustered to the right, along CCA-1, while the different label categories are separable from top to bottom, along CCA-2. Ultimately, visualizations in Figure 3 reflect prior interpretations of the CCA R-values and weights: dimension 1 primarily reflects prosodic emphasis of syllables (strong beats with high-or-low pitch tones) and dimension 2 reflects pitch tone level (high versus low or combined high and low).

The distributions in Figure 3 indicate that different RaP label conjunctions are highly discriminable from each other. The next goal is to predict RaP label conjunctions of interest from the acoustic features. For this, a Gaussian Classifier (GC) was trained to detect certain label conjunctions. Specifically, three GCs were trained, to detect the following: 1. nXx & (E or nT); 2. (x or X) & (L or H or H&L); 3. X & (L or H or H&L). These categories respectively represent a scale ranging from syllables with no beats and no tones to those with strong metrical beats and significant pitch variability (high or low tones), effectively mapping out the dominant relationship revealed by CCA. The discrimination evaluations were done using leave-one-out cross-validation, with the GC output (two-class log-likelihood ratios) used to compute receiver operating characteristic (ROC) curves. Area under the ROC curve (AUC) values of 0.78, 0.80, and 0.84 were obtained for the three label conjunctions, respectively. These results indicate a strong ability to detect, solely from acoustic features, dominant prosodic content, particularly emphasized syllables representing strong beats with significant pitch variability.

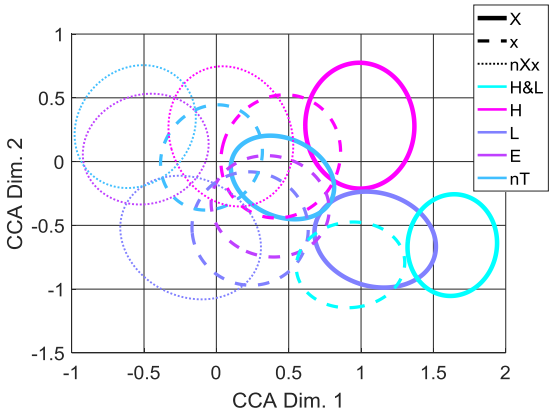

Figure 3: *Categorical RaP clusters across CCA-1 and CCA-2 feature dimensions. Rhythm key:X-solid, x-dashed, nXx-dotted.*

5. Conclusions & Discussion

Exploiting recent linguistic advances in Rhythm and Pitch (RaP) transcription, the present work provides a statistical framework for analysis and prediction of prosodic events from speech acoustics. Canonical Correlation Analysis (CCA) reveals underlying relationships between syllable durational, fundamental frequency and intensity acoustic features and perceived annotated rhythmic prominences (metrical beats) and pitch tones. In particular, the dominant correlation component reveals a strong relationship between prosodically emphasized syllables (strong beats with high or low pitch tones) indicated primarily by duration, followed by mean intensity and fundamental frequency. This finding motivates revisiting common HMM-based phone durational models used in speech recognition and synthesis contexts [23]. The aggregated treatment of phones in these models does not capture durational variation due to rhythmic patterning observed in natural speech. The second primary axis of correlation discriminates the pitch tone level, as also demonstrated in the prediction results. Finally, good accuracy and generalizability were achieved in predicting the prosodically emphasized versus unstressed (no beat, no tone) syllables, ultimately suggesting that automatic acoustic analyses present a future capability for high-fidelity prosodic labelling that could impact a range of speech technologies.

6. Acknowledgements

Thank you to Dr. Laura Dilley at Michigan State University and Dr. Maura Breen at Mount Holyoke College for sharing the RaP annotated corpora used in this work.

7. References

[1] M. Wagner, and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and cognitive processes* vol. 25.no. 7-9, pp. 905-945, 2010.

[2] R. Fernandez, A. Rosenberg, A. Sorin, B. Ramabhadran, and R. Hoory, "Voice-Transformation-Based Data Augmentation for Prosodic Classification," in *ICASSP*, pp. 5530—5534, 2017.

[3] L. Dilley, and M. Brown, "The RaP (rhythm and pitch) labeling system. v. 1.0," *Massachusetts Institute of Technology*, 2005.

[4] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody." In *Proceedings of the 2nd International Conference on Spoken Language Processing,* pp. 867-879, 1992.

[5] M. Breen, L. Dilley, J. Kraemer, and E. Gibson, "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)," *Corpus Linguistics and Linguistic Theory*, vol. 8, no. 2, pp. 277-312, 2012.

[6] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[7] S. Ananthakrishnan, and S. Narayanan, "Fine-grained pitch accent and boundary tone labeling with parametric F0 features," in *ICASSP*, pp. 4545–4549, 2008.

[8] H. Pfitzinger, "Local Speech Rate as a Combination of Syllable and Phone Rate," in *ICSLP*, 1998.

[9] V. Dellwo, "The role of speech rate in perceiving speech rhythm," *Speech Prosody*, pp. 375–378, 2008.

[10] R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," *Journal of Phonetics*, vol. 11, pp. 51–62, 1983.

[11] R. M. Dauer, "Phonetic and phonological components of language rhythm," *International Congress of Phonetic Sciences*, pp. 447–450, 1987.

[12] P. Martin, "Prominence Detection without Syllabic Segmentation," *Proceedings of Prosodic Prominence, Speech Prosody,* 2010.

[13] W. Härdle and L. Simar, "Canonical Correlation Analysis," *Applied Multivariate Statistical Analysis*. pp. 321–330, 2007.

[14] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation MIT* 16, pp. 2639–2664, 2004.

[15] N. Correa, Y. O. Li, T. Adali, and V. D. Calhoun, "Canonical Correlation Analysis for Feature-Based Fusion of Biomedical Imaging Modalities and Its Application to Detection of Associative Networks in Schizophrenia," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 998—1007, 2008.

[16] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech LDC97S42 DVD," *Philadelphia: Linguistic Data Consortium*, 1997.

[17] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "Boston University Radio Speech Corpus LDC96S36. Web Download," *Philadelphia: Linguistic Data Consortium*, 1996.

[18] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in Laboratory Phonology*, vol. 7, pp. 515–546, 2002.

[19] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International,* vol. 5, no. 9/10, pp. 341-345, 2001.

[20] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, T. S. Huang, "Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction," in *International Workshop on Audio/Visual Emotion Challenge*, pp. 97-104, 2016.

[21] R. L. Horwitz-Martin, T. F Quatieri, E. Godoy, J. R. Williamson, "A vocal modulation model with application to predicting depression severity," in *Body Sensor Networks*, pp. 247-25, 2016.

[22] A. Röbel and X. Rodet, "Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation," *DAF'x'05 Spain*, 2005.

[23] The HTK Book. Cambridge University, 2002.