Interspeech 2018
2-6 September 2018, Hyderabad

Towards an Unsupervised Entrainment Distance in Conversational Speech using Deep Neural Networks

Md Nasir[1], *Brian Baucom*[2], *Shrikanth Narayanan*[1], *Panayiotis Georgiou*[1]

[1]University of Southern California, Los Angeles, CA, USA
[2]University of Utah, Salt Lake City, UT, USA

mdnasir@usc.edu, brian.baucom@utah.edu, {shri, georgiou}@sipi.usc.edu

Abstract

Entrainment is a known adaptation mechanism that causes interaction participants to adapt or synchronize their acoustic characteristics. Understanding how interlocutors tend to adapt to each other's speaking style through entrainment involves measuring a range of acoustic features and comparing those via multiple signal comparison methods. In this work, we present a turn-level distance measure obtained in an unsupervised manner using a Deep Neural Network (DNN) model, which we call *Neural Entrainment Distance* (NED). This metric establishes a framework that learns an embedding from the population-wide entrainment in an unlabeled training corpus. We use the framework for a set of acoustic features and validate the measure experimentally by showing its efficacy in distinguishing real conversations from fake ones created by randomly shuffling speaker turns. Moreover, we show real world evidence of the validity of the proposed measure. We find that high value of NED is associated with high ratings of emotional bond in suicide assessment interviews, which is consistent with prior studies.

Index Terms: entrainment, deep neural network, unsupervised learning, embeddings, behavioral analysis, conversational speech

1. Introduction

Vocal entrainment is an established social adaptation mechanism. It can be loosely defined as one speaker's spontaneous adaptation to the speaking style of the other speaker. Entrainment is a fairly complex multifaceted process and closely associated with many other mechanisms such as coordination, synchrony, convergence *etc.* While there are various aspects and levels of entrainment [1], there is also a general agreement that entrainment is a sign of positive behavior towards the other speaker [2–4]. High degree of vocal entrainment has been associated with various interpersonal behavioral attributes, such as high empathy [5], more agreement and less blame towards the partner and positive outcomes in couple therapy [6], and high emotional bond [7]. A good understanding of entrainment provides insights to various interpersonal behaviors and facilitates the recognition and estimation of these behaviors in the realm of Behavioral Signal Processing [8, 9]. Moreover, it also contributes to the modeling and development of 'human-like' spoken dialog systems or conversational agents.

Unfortunately, quantifying entrainment has always been a challenging problem. There is a scarcity of reliable labeled speech databases on entrainment, possibly due to the subjective and diverse nature of its definition. This makes it difficult to capture entrainment using supervised models, unlike many other behaviors. Early studies on entrainment relied on highly subjective and context-dependent manual observation coding for measuring entrainment. The objective methods based on extracted speech features employed classical synchrony measures such as Pearson's correlation [1] and traditional (linear) time series analysis techniques [10]. Lee *et al.* [5, 11] proposed a measure based on PCA representation of prosody and MFCC features of consecutive turns. Most of these approaches assume a linear relationship between features of consecutive speaker turns which is not necessarily true, given the complex nature of entrainment. For example, the effect of rising pitch or energy can potentially have a nonlinear influence across speakers.

Recently, various complexity measures (such as largest Lyapunov exponent) of feature streams based on nonlinear dynamical systems modeling showed promising results in capturing entrainment [6, 7]. A limitation of this modeling, however, is the assumption of the short-term stationary or slowly varying nature of the features. While this can be reasonable for global or session-level complexity, the measure is not very meaningful capturing turn-level or local entrainment. Nonlinear dynamical measures also suffer from scalability to a multidimensional feature set, including spectral coefficients such as MFCCs. Further, all of the above metrics are knowledge-driven and do not exploit the vast amount of information that can be gained from existing interactions.

A more holistic approach is to capture entrainment in consecutive speaker turns through a more robust nonlinear function. Conceptually speaking, such a formulation of entrainment is closely related to the problem of learning a transfer function which maps vocal patterns of one speaker turn to the next. A compelling choice to nonlinearly approximate the transfer function would be to employ Deep Neural Networks (DNNs). This is supported by recent promising applications of deep learning models, both in supervised and unsupervised paradigms, in modeling and classification of emotions and behaviors from speech. For example in [12] the authors learned, in an unsupervised manner, a latent embedding towards identifying behavior in out-of-domain tasks. Similarly in [13, 14] the authors employ Neural Predictive Coding to derive embeddings that link to speaker characteristics in an unsupervised manner.

We propose an *unsupervised training* framework to *contextually* learn the transfer function that ties the two speakers. The learned bottleneck embedding contains cross-speaker information closely related to entrainment. We define a distance measure between the consecutive speaker turns represented in the bottleneck feature embedding space. We call this metric the ***Neural Entrainment Distance*** **(NED)**.

Towards this modeling approach we use features that have already been established as useful for entrainment. The majority of research [1, 6, 7, 11, 15] focused on prosodic features like

10.21437/Interspeech.2018-1395

pitch, energy, and speech rate. Others also analyzed entrainment in spectral and voice quality features [5, 11]. Unlike classical nonlinear measures, we jointly learn from a *multidimensional feature set* comprising of prosodic, spectral, and voice quality features.

We then experimentally investigate the validity and effectiveness of the NED measure in association with interpersonal behavior.

2. Datasets

We use two datasets in this work: the training is done on the Fisher Corpus English Part 1 (LDC2004S13) [16] and testing on the Suicide Risk Assessment corpus [17], along with Fisher.

- **Fisher Corpus English Part 1**: It consists of spontaneous telephonic dyadic conversations between native English speakers. There are 5850 such conversations, each lasting up to 10 minutes. The manual transcripts of the corpus contain time-stamps of speaker turn boundaries as well as boundaries of pauses within a turn.
- **Suicide Risk Assessment corpus**: This dataset contains recorded conversations of active duty military personnel with their therapist during suicide risk assessment sessions. The participants were suicidal patients who either had attempted suicide or had suicidal thoughts prior to the sessions. The subset of the corpus employed in the current work included therapist-patient interviews of 54 subjects, each session ranging from 10 minutes to 1 hour. They were asked questions related to their personal history, reasons leading to their suicidal ideations, elaboration of their reasons for living *etc.* Immediately after the interview sessions, the patient was asked to provide with a self-reported score for perceived *emotional bond*, an attribute which entails the therapist's empathy for the patient and the patient's feeling of trust towards them. It was rated on a scale from 1 to 10.

3. Modeling of Neural Entrainment Distance

3.1. Preprocessing

A number of audio preprocessing steps are required in the entrainment framework for obtaining boundaries of relevant segments of audio from consecutive turns. First, we perform voice activity detection (VAD) to identify the speech regions. Following this, speaker diarization is performed in order to distinguish speech segments spoken by different speakers. However, our training dataset, the Fisher corpus also contains transcripts with speaker turn boundaries as well as timings for pauses within a turn. Since, these time stamps appeared to be reasonably accurate, we use them as oracle VAD and diarization. On the other hand, for the Suicide Risk Assessment corpus, we perform VAD and diarization on raw audio to obtain the turn boundaries. Subsequently, we also split a single turn into inter-pausal units (IPUs) if there is any pause of at least 50 ms present within the turn. For the purpose of capturing entrainment-related information, we only consider the initial and the final IPU of every turn. This is done based on the hypothesis that during a turn-taking, entrainment is mostly prominent between the most recent IPU of previous speaker's turn and the first IPU of the next speaker's turn [1].

3.2. Feature Extraction

We extract 38 different acoustic features from the segments (IPUs) of our interest. The extracted feature set includes 4 prosody features (pitch, energy and their first order deltas), 31 spectral features (15 MFCCs, 8 MFBs, 8 LSFs) and 3 voice quality features (shimmer and 2 variants of jitter). We found in our early analysis that derivatives of spectral and voice quality features do not seem to contribute significantly to entrainment[1] and hence we do include them for the NED model. The feature extraction is performed with a Hamming window of 25 ms width and 10 ms shift using the OpenSMILE toolkit [18]. For pitch, we perform an additional post-processing by applying a median-filter based smoothing technique (with a window size of 5 frames) as pitch extraction is not very robust and often prone to errors, such as halving or doubling errors. We also perform z-score normalization of the features across the whole session, except for pitch and energy features, which are normalized by dividing them by their respective means.

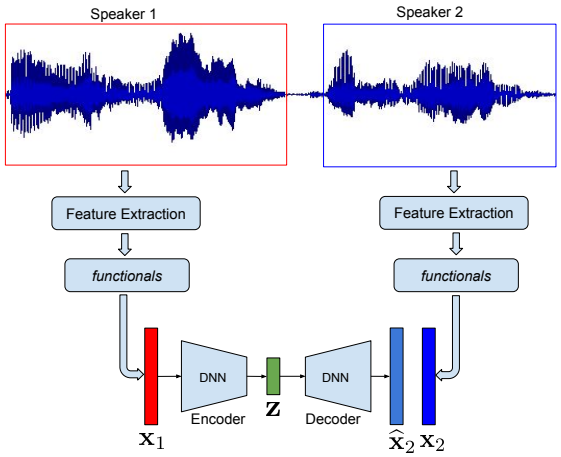

Figure 1: *An overview of unsupervised training of the model*

3.3. Turn-level Features

We propose to calculate NED as directional entrainment-related measure from speaker 1 to speaker 2 for a change of turn as shown in Figure 1. The segments of interest in this case are the final IPU of speaker 1's turn and the initial IPU of the subsequent turn by speaker 2, marked by the bounding boxes in the figure. As turn-level features, we compute six statistical functionals over all frames in those two IPUs, generating two sets of functionals of features for each pair of turns. The functionals we compute are as follows: mean, median, standard deviation, 1st percentile, 99th percentile and range between 99th and 1st percentile. Thus we obtain $38 \times 6 = 228$ turn-level features from each IPU representing the turn. Let us denote the turn-level feature vector of the final IPU of speaker 1 and the initial IPU of speaker 2 as $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, for further discussion in the paper.

3.4. Modeling with Neural Network

Most work in the entrainment literature directly computes a measure between $\mathbf{x}_1$ and $\mathbf{x}_2$ (such as correlation [1]) or their lower-dimensional representations [11]. However, one conceptual limitation of all these approaches is that turn-level

[1]These features showed very low correlation ($\rho < 0.05$) across consecutive turns in our initial analysis

features $\mathbf{x}_1$ and $\mathbf{x}_2$ do not only contain the underlying acoustic information that can be entrained across turns, but also speaker-specific, phonetic and paralinguistic information that is specific to the corresponding turns and not influenced by the previous turn (non-entrainable). If we represent those two types of information as vector embeddings, **e** and **q** respectively, we can model turn-level feature vectors **x** as a nonlinear function $\mathscr{F}(\cdot)$ over them, *i.e.,* $\mathbf{x}_1 = \mathscr{F}(\mathbf{e}_1, \mathbf{q}_1)$ and $\mathbf{x}_2 = \mathscr{F}(\mathbf{e}_2, \mathbf{q}_2)$. In this formulation, the distance between $\mathbf{e}_1$ and $\mathbf{e}_2$ should be zero in the hypothetical case of 'perfect' entrainment.

Our goal is to approximate the inverse mappings that maps the feature vector **x** to entrainment embedding **e** and ideally to learn the same from 'perfect' or very highly entrained turns. Unfortunately, in absence of such a dataset, we learn it from consecutive turns in real data where entrainment is present, at least to some extent. As shown in Figure 1, we adopt a feed-forward deep neural network (DNN) as an encoder for this purpose.

The different components of the model are described below:

1. First we use $\mathbf{x}_1$ as the input to the encoder network. We choose the output of the encoder network, **z** to be undercomplete representation of $\mathbf{x}_1$, by restricting the dimensionality of **z** to be lower than that of **x**.
2. **z** is then passed through another feed-forward (**z**) network used as decoder to predict $\mathbf{x}_2$. The output of the decoder is denoted as $\widehat{\mathbf{x}}_2$.
3. Then $\widehat{\mathbf{x}}_2$ and its reference $\mathbf{x}_2$ are compared to obtain the loss function of the model, $\mathcal{L}(\mathbf{x}_2, \widehat{\mathbf{x}}_2)$.

Even though this deep neural network resembles autoencoder architectures, it does not reconstruct itself but rather tries to encode relevant information from one turn to predict the next turn, parallel to [12–14]. Thus the bottleneck embedding **z** can be considered closely related to the entrainment embedding **e** mentioned above.

3.5. Unsupervised Training of the Model

In this work, we use two fully connected layers as hidden layers both in the encoder and decoder network. Batch normalization layers and Rectified Linear Unit (ReLU) activation layers (in respective order) are used between fully connected layers in both of the networks. The dimension of the embedding is chosen to be 30. The number of neuron units in the hidden layers are: [228 → 128 → 30 → 128 → 228]. We use smooth L1 norm, a variant of L1 norm which is more robust to outliers [19], so that

$$\mathcal{L}(\mathbf{x}_2, \widehat{\mathbf{x}}_2) = \|\mathbf{x}_2 - \widehat{\mathbf{x}}_2\|_1^{\text{smooth}} = \sum_{k=1}^{N} \text{smooth}_{L_1}(x_{2k} - \widehat{x}_{2k}), \quad (1)$$

where

$$\text{smooth}_{L_1}(d) = \begin{cases} 0.5d^2, & \text{if } |d| \leq 1 \\ |d| - 0.5, & \text{otherwise} \end{cases} \quad (2)$$

and N is the dimension of **x** which is 228 in our case.
For training the network, we choose a subset (80% of all sessions) of Fisher corpus and use all turn-level feature pairs $(\mathbf{x}_1, \mathbf{x}_2)$. We employ the Adam optimizer [20] and a minibatch size of 128 for training the network. The validation error is computed on the validation subset (10% of the data) of the Fisher corpus and the best model is chosen.

3.6. Neural Entrainment Distance (NED) Measure

After the unsupervised training phase, we use the encoder network to obtain the embedding representation (**z**) from any turn-level feature vector **x**. To quantify the entrainment from a turn to the subsequent turn, we extract turn-level feature vectors from their final and initial IPUs, respectively, denoted as $\mathbf{x}_i$ and $\mathbf{x}_j$. Next we encode $\mathbf{x}_i$ and $\mathbf{x}_j$ using the pretrained encoder network and obtain $\mathbf{z}_i$ and $\mathbf{z}_j$ as the outputs, respectively. Then we compute a distance measure d_{NE}, which we term *Neural Entrainment Distance* (NED), between the two turns by taking smooth L1 distance $\mathbf{z}_i$ and $\mathbf{z}_j$.

$$d_{\text{NE}}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{z}_i - \mathbf{z}_j\|_1^{\text{smooth}} = \sum_{k=1}^{M} \text{smooth}_{L_1}(z_{ik} - z_{jk}), \quad (3)$$

where $\text{smooth}_{L_1}(\cdot)$ is defined in Equation (2) and M is the dimensionality of the embedding. Note that even though smooth L1 distance is symmetric in nature, our distance measure is still asymmetric because of the directionality in the training of the neural network model.

4. Experimental Results

We conduct a number of experiments to validate NED as a valid proxy metric for entrainment.

4.1. Experiment 1: Classification of real *vs.* fake sessions

We first create a fake session ($\mathcal{S}_{fake}$) from each real session ($\mathcal{S}_{real}$) by randomly shuffling the speaker turns. Then we run a simple classification experiment of using the NED measure to identify the real session from the pair ($\mathcal{S}_{real}, \mathcal{S}_{fake}$). The steps of the experiments are as follows:

1. We compute NED for each (overlapping) pair of consecutive turns and their average across the session for both sessions in the pair ($\mathcal{S}_{real}, \mathcal{S}_{fake}$).
2. The session with lower NED is inferred to be the real one. The hypothesis behind this rule is that higher entrainment is seen across consecutive turns than randomly paired turns and is well captured through a lower value of proposed measure.
3. If the inferred real session is indeed the real one, we consider it to be correctly classified.

We compute classification accuracy averaged over 30 runs (to account for the randomness in creating the fake session) and report it in Table 1. The experiment is conducted on two datasets: a subset (10%) of Fisher corpus set aside as test data and Suicide corpus. We use a number of baseline measures:

- Baseline 1: smooth L1 distance directly computed between turn-level features ($\mathbf{x}_i$ and $\mathbf{x}_j$)
- Baseline 2: PCA-based symmetric acoustic similarity measure by Lee *et al.* [11]
- Baseline 3: Nonlinear dynamical systems-based complexity measure [7].

For the baselines, we conduct the classification experiments in a similar manner. Since Baseline 1 and 2 have multiple measures, we choose the best performing one for reporting, thus providing an upper-bound performance. Also, for baseline 2 we choose the session with higher value of the measure as real, since it measures similarity.

As we can see in Table 1, our proposed NED measure achieves higher accuracy than all baselines on the Fisher corpus. The accuracy of our measure declines in the Suicide corpus

Measure	Classification accuracy (%)	
	Fisher corpus	Suicide corpus
Baseline 1	72.10 (5.83)	70.44 (6.69)
Baseline 2	92.32 (3.01)	88.12 (5.93)
Baseline 3	90.21 (5.40)	88.54 (5.87)
NED	98.87 (0.97)	91.92 (2.32)

Table 1: *Results of Experiment 1: classification accuracy (%) of real vs. fake sessions (averaged over 30 runs; standard deviation shown in parentheses)*

as compared to the Fisher corpus, which is probably due to data mismatch as the model was trained on Fisher (mismatch of acoustics, recording conditions, sampling frequency, interaction style *etc.*). However, our measure still performs better than all baselines on Suicide corpus.

4.2. Experiment 2: Correlation with Emotional Bond

According to prior work, both from domain theory [17] and from experimental validation [7], a high emotional bond in patient-therapist interactions in the suicide therapy domain is associated with more entrainment. In this experiment, we compute the correlation of the proposed NED measure with the patient-perceived emotional bond ratings. Since the proposed measure is asymmetric in nature, we compute the measures for both patient-to-therapist and therapist-to-patient entrainment. We also compute the correlation of emotional bond with the baselines used in Experiment 1. We report Pearson's correlation coefficients (ρ) for this experiment in Table 2 along with their p-values. We test against the null hypothesis H_0 that there is no linear association between emotional bond and the candidate measure.

Results in Table 2 show that the patient-to-therapist NED is negatively correlated with emotional bond with high statistical significance ($p < 0.01$). This negative sign is consistent with previous studies as higher distance in acoustic features indicates lower entrainment. However, the therapist-to-patient NED does not have a significant correlation with emotional bond. A possible explanation for this finding is that the emotional bond is reported by the patient and influenced by the degree of their perceived therapist-entrainment. Thus, equipped with an asymmetric measure, we are also able to identify the latent directionality of the emotional bond metric. The complexity measure (Baseline 2) also shows statistically significant correlation, but the value of ρ is lower than that of the proposed measure.

To analyze the embeddings encoded by our model, we also compute a t-SNE [21] transformation of the difference of all

Measure	Pearson's correlation	
	ρ	p-value*
Baseline 1	-0.1980	0.2031
Baseline 2	0.2480	0.1132
Baseline 3	-0.3815	0.0127
NED-TP	-0.1317	0.3999
NED-PT	-0.4479	0.0095

Table 2: *Correlation between emotional bond and various measures; TP: therapist-to-patient, PT: patient-to-therapist*

*$p < 0.05$ *indicates statistically significant correlation*

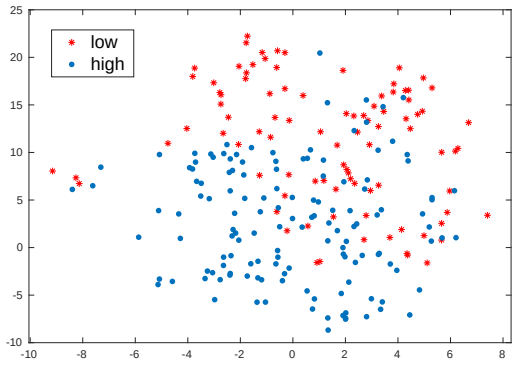

Figure 2: *t-SNE plot of difference vector of encoded turn-level embeddings for sessions with low and high emotional bond*

patient-to-therapist turn embedding pairs, denoted as $\mathbf{z}_i - \mathbf{z}_j$ in Equation (3). Figure 2 shows the results of a session with high emotional bond and another one with low emotional bond (with values of 7 and 1 respectively) as a 2-dimensional scatter plot. Visibly there is some separation between the sessions with low and high emotional bond.

5. Conclusion and Future Work

In this work, a novel deep neural network-based *Neural Entrainment Distance* (NED) measure is proposed for capturing entrainment in conversational speech. The neural network architecture consisting of an encoder and a decoder is trained on the Fisher corpus in an unsupervised training framework and then the measure is defined on the bottleneck embedding. We show that the proposed measure can distinguish between real and fake sessions by capturing presence of entrainment in real sessions. In this way we also validate the natural occurrence of vocal entrainment in dyadic conversations, well-known in psychology literature [22–24]. We further show that the measure for patient-to-therapist direction achieves statistically significant correlation with their perceived emotional bond. The proposed measure is asymmetric in nature and can be useful for analyzing different interpersonal (especially directional) behaviors in many other applications. Given the benefits shown by the unsupervised data-driven approach we will employ Recurrent Neural Networks (RNNs) to better capture temporal dynamics. We also intend to explore (weakly) supervised learning of entrainment using the bottleneck embeddings as features, in presence of session-level annotations.

6. Acknowledgements

The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702- 5014 is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Military Suicide Research Consortium under Award No. W81XWH-10-2-0181, and through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

7. References

[1] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[2] J. Welkowitz and M. Kuc, "Interrelationships among warmth, genuineness, empathy, and temporal speech patterns in interpersonal interaction." *Journal of Consulting and Clinical Psychology*, vol. 41, no. 3, p. 472, 1973.

[3] J. Hirschberg, "Speaking more like you: Entrainment in conversational speech," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[4] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, no. 4, pp. 802–813, 2014.

[5] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling." in *INTERSPEECH*, 2013, pp. 2861–2865.

[6] M. Nasir, B. Baucom, S. S. Narayanan, and P. Georgiou, "Complexity in prosody: A nonlinear dynamical systems approach for dyadic conversations; behavior and outcomes in couples therapy," *Interspeech 2016*, pp. 893–897, 2016.

[7] M. Nasir, B. Baucom, C. J. Bryan, S. Narayanan, and P. Georgiou, "Complexity in speech and its relation to emotional bond in therapist-patient interactions during suicide risk assessment interviews," in *Proceedings of Interspeech. August 2017*, 2017.

[8] P. G. Georgiou, M. P. Black, and S. S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. Scottsdale, AZ: ACM, 2011, pp. 7–12.

[9] S. Narayanan and P. G. Georgiou, "Behavioral Signal Processing: deriving human behavioral informatics from speech and language," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 101, no. 5, p. 1203, 2013.

[10] S. Kousidis, D. Dorran, C. Mcdonnell, and E. Coyle, "Convergence in human dialogues time series analysis of acoustic feature," 2009.

[11] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[12] H. Li, B. Baucom, and P. Georgiou, "Unsupervised latent behavior manifold learning from acoustic features: Audio2behavior," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, March 2017.

[13] A. Jati and P. Georgiou, "Neural predictive coding using convolutional neural networks towards unsupervised learning of speaker characteristics," *IEEE Trans. Speech, Audio, and Language Processing*, 2018.

[14] ——, "Speaker2vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation," in *Proceedings of Interspeech*, Stockholm, Sweden, August 2017.

[15] C.-C. Lee, M. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[16] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[17] B. Baucom, A. Crenshaw, C. Bryan, T. Clemans, T. Bruce, and M. Rudd, "Patient and clinician vocally encoded emotional arousal as predictors of response to brief interventions for suicidality," *Brief Cognitive Behavioral Interventions to Reduce Suicide Attempts in Military Personnel. Association for Behavioral and Cognitive Therapies*, 2014.

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

[19] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1440–1448. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.169

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] J. H. Watt and C. A. VanLear, *Dynamic patterns in communication processes*. Sage Publications, Inc, 1996.

[23] P. A. Andersen and J. F. Andersen, "The exchange of nonverbal intimacy: A critical review of dyadic models," *Journal of Nonverbal Behavior*, vol. 8, no. 4, pp. 327–349, 1984.

[24] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 2007.