MULTISTREAM DIARIZATION FUSION USING THE MINIMUM VARIANCE BAYESIAN INFORMATION CRITERION

Tae Jin Park, Panayiotis Georgiou

University of Southern California, Los Angeles, CA, USA

taejinpa@usc.edu, georgiou@sipi.usc.edu

ABSTRACT

Speaker diarization is necessary with ubiquitous and individualized recorders. We focus on the specific task of speaker diarization from two information streams, two microphones, assigned to two participants of interest. In real scenarios, speakers may be co-located, in noisy environments with interfering speakers. Multistream diarization can exploit additional information and diarization fusion is necessary. In this work we first introduce a new database that realistically simulates a range of extremely challenging acoustic conditions; and propose a Minimum Variance of BIC (MVBIC) method to combine information from the various diarization streams. We use a 2-microphone subset of our proposed database and Root Mean Square Energy (RMSE) and Mel Frequency Cepstral Coefficients (MFCC) as our two diarization streams to validate the proposed method. We show that our proposed method exploits the complementarity of the individual diarization streams and outperforms static fusion mixing weights. We also demonstrate the robustness of the MVBIC method on RT-06S data.

Index Terms— Speaker Diarization, Diarization Fusion, Individual Near-field Microphone, Bayesian Information Criterion

1. INTRODUCTION

Significant work has taken place over the last decade in speaker diarization. Multiple diarization challenges and datasets have been created and extensive research has taken place in diarization from a single microphone. NIST challenges have also included Single Distant Microphone (SDM) and Multiple Distant Microphones (MDM) [1, 2] diarization.

One of the domains of significant interest recently has been analyzing human behavior. Our group has been very active in the field with work in several Behavioral Signal Processing [3, 4] domains such as in Addiction [5, 6], Couples Therapy [7, 8], Suicide [9], Cancer [10], *etc.*. In all of these domains we are increasingly seeing signals very different from all the existing available diarization corpora. While internally we can annotate and evaluate our algorithms on such data, it is difficult to share these due to privacy concerns.

The main differences we observe from available corpora and from the real-world data we see involve the quality of signals, the high variability conditions, the availability of multiple microphones often one or more assigned to a specific individual, and the presence of interfering sources. One example scenario is when two individuals have their own individual recording devices and record their interaction throughout multiple days while they are at home. In such a scenario, which we will call for the sake of convenience the Multiple Individualized (near-field) Microphones (MIM) condition, we often have visitors or TV in the environment that cause interference, but also beneficial additional information due to the microphone placement. For example we can have multiple diarization streams from the multiple microphones, diarization from the mic relative pickup energy, diarization from the Time Difference Of Arrival (TDOA) *etc.*.

In this work, we will first present an overview of the creation of a challenging diarization corpus that closely simulates the real world environments presented above. Then we will tackle the task of how we can robustly fuse multiple diarization schemes. As a first task in this paper, we will use the subset of the presented corpus that contains only 2 speakers and employ only 2 streams for the diarization fusion. This is not limiting the scope of the work but focuses on a narrower scenario, which closer resembles our data, as a first stage.

2. THE CREATION OF A DATASET

We created the USCDiarLibri dataset that can be used to test speaker diarization tasks with various customized setups and randomization. The creation protocol is open-source and available to the public from our website `http://scuba.usc.edu/software`. It is based on artificial multi-party dialogs made from noisy, reverberated audio from the LibriSpeech [11] database and it's highly parameterized to allow for diverse conditions. In this paper we describe the small part of the corpus used for the study, USCDiarLibri2,4, while a more extensive description of the USCDiarLibri corpus is in preparation and will also be on our git repo.

2.1. Dialog in USCDiarLibri2,4

USCDiarLibri2,4 consists of 130 sessions (80 used in Sec.4.1 and 50 used in Sec. 4.2). Each session includes 2-channel recordings from 4 unique speakers. Two of those speakers are the speakers of interest while the others are interfering.

The speech data comes from the LibriSpeech ASR corpus [11]. We employed force-alignment to extract accurate word boundaries. We generated the number of words in each speaker turn with Rayleigh random variable with mean of 7.5, which reflects the skewed distribution of speech segment duration [12]. The total running time of each session varies due to the available amount of speech data from each of the 4 speakers in the original LibriSpeech corpus, and ranges from 7min to 20min.

To generate the dialog the system goes through 6 states: Speakers 1 through 4 (S_i) with probability $p = 0.2$, No Speech (N_S) with $p = 0.1$, and Overlap (O) also with $p = 0.1$.

• States S_1 through S_4 are inserting a new turn by the corresponding speaker right after the previous turn.

• State N_s introduces a gap between turns, random in duration $[0, 5\text{sec}]$.

• The O state introduces two turns, one by each the main speakers S_1 and S_2, that are overlapping. One of the two speakers is picked at random as the first overlap-segment. The other speaker can be activated at any time during the time the first speaker is active (uniform probability). The O state ends when both S_1 and S_2 turns are completed.

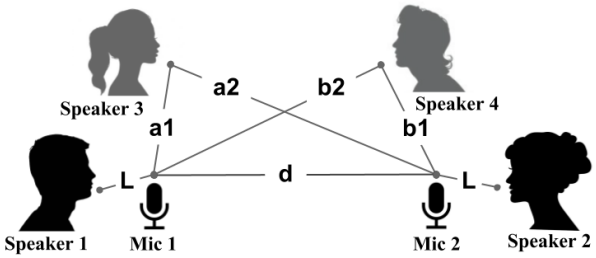

Fig. 1. USCDiarLibri2,4 assumes 2 speakers of interest among 4 active speakers. It models reverberation, overlap, and interfering sources.

2.2. Impulse Response and Spatial Simulation

Figure 1 describes the dimensional information of USCDiarLibri2,4 dataset. Two primary speakers talk to their microphones (Speaker 1 and Speaker 2 in Fig 1) and the other two interfering speakers (Speaker 3 and Speaker 4 in Fig 1) are co-located around primary speakers. We assume that each of the primary speakers has their own microphone at a short distance L=0.3m from their mouth. The distances between speakers are fixed during a session. To simulate acoustical degradation of distant speech in real life, we employed simulated Impulse Response (IR) [13]. We also set absorption coefficient as 0.25 to simulate an echoic room. The amplitude and time delay of speech signal is simulated according to distance between a speaker and a microphone as below:

$$x_{\text{mic}}[t] = \frac{1}{r}\,\bar{H}[t] * x_{\text{source}}\left[t - r\frac{f_s}{v_s}\right] \tag{1}$$

where x_{mic} is signal picked up from the microphone, r is the distance between the speaker and the microphone, f_s is the sampling rate, v_s is the speed of sound and $\bar{H}$ is the amplitude normalized impulse response between the speaker and the microphone that filters the original source signal x_{source}. The dimension of the virtual echoic room is determined according to distances between speakers, and IR is convoluted with the source signal.

As mentioned above, due to space limitations, we only describe the limited subset of USCDiarLibri that is employed in the next section. We want to highlight that the full USCDiarLibri encompases more realistic turntaking, higher control of overlapping sources, controllable number of sources of interest and of interference. For the scope of this work the subset USCDiarLibri2,4 is sufficient.

3. PROPOSED DIARIZATION FUSION FRAMEWORK

The diarization task for two individual microphones, when signals contain no noise or interference, can be achieved with near perfect accuracy by employing power level differences [14]. However, in a real scenario, where people are going about their daily lives, interacting with third parties, and under variable acoustic conditions and relative locations, the diarization task becomes more difficult. It's not uncommon for instance for the "other" microphone to be picking up source louder than the "own" microphone, or both microphones to be picking up a third speaker.

In such more challenging scenarios, with multiple recordings, using multiple diarization schemes can be beneficial. For example, we can have diarization using various features such as Mel Frequency Cepstrum Coefficients (MFCC), TDOA or relative signal energy. However the fusion of those diarization decisions becomes more challenging due to the variable reliability of each diarization algorithm over different acoustic conditions.

Solely relying on one information does not give the best performance [15] because in each diarization task case, speaker characteristics, distances between speakers and room impulse responses all play roles and make an unpredictable environment. Since MDM condition also suffers from this issue, the authors in [15, 16] approached this problem by determining a fixed weight factor by optimizing on a dev-set to make a compound likelihood model (BIC) based on MFCC and TDOA.

In real world data however, people interact in diverse conditions and fusion with pre-determined weights from a dev set leads in extreme performance drop. We therefore require, and propose, a Minimum Variance of Bayesian Information Criterion (MVBIC) to effectively combine the various diarization streams. The proposed MVBIC technique does not fix a weight factor over sessions as in the previous work [15, 16] but estimates an effective weight for each individual session[1].

To generate our two sample diarization streams, we employed the Root Mean Square Energy (RMSE) and MFCC features. We chose RMSE over the TDOA feature used in RT-06S MDM condition due to the significant drop of performance of TDOA under reverberant conditions with interfering speakers as described in [17].

3.1. Unimodal Speaker Diarization

Our efforts in this work are to improve multi-diarization fusion. As such, our individual stream diarization is employing the most widely known algorithms without any modifications. We employ the prototypical diarization frame work proposed in [18]. This framework comprises Speech Activity detection (SAD), Segmentation, and bottom-up clustering algorithm based on Bayesian Information Criterion (BIC):

• *SAD:* we employed the widely used algorithm provided by [19].

• *Speech Segmentation:* we employed segmentation technique based on KL distance in [20]. For all the diarization work in this paper, we use sum of two KL distances from each of two feature streams to control the effect of segmentation performance. We used a window length of 50.

• *Clustering:* we employed the formula for BIC and the clustering method described in [21].

In this work, we use two diarization feature streams: One is based on a two-dimensional RMSE feature [22] from each channel and the second on a 13-dimensional MFCC feature [23] extracted from mean of two channels.

3.2. Diarization Fusion: Minimum Variance BIC

We propose the Minimum Variance BIC (MVBIC) technique that efficiently weights BIC distances according to their reliability towards improved clustering accuracy. The concept of minimum variance optimization has also appeared in the studies from other fields, such as finance [24] or acoustics [25].

We assume that there is an underlying correct BIC stream that we are observing through a noisy channel. The hidden, correct BIC stream will be represented by b and its two observed, noisy versions by $\tilde{b}_i$, where in our case $i \in [1, M]$ and $M = 2$ (MFCC and RMSE). Therefore:

$$\tilde{b}_i = b + n_i \tag{2}$$

where the above three are all random variables.

With the above model (2), we want to obtain the optimal fusion weights that will lead to accurate estimation of the true b value:

$$\widehat{b} = \sum_{i=1}^{M} \omega_i\, b_i = \mathbf{w}^T\mathbf{b} \tag{3}$$

[1] In a real world scenario, this can be a sliding window of several minutes.

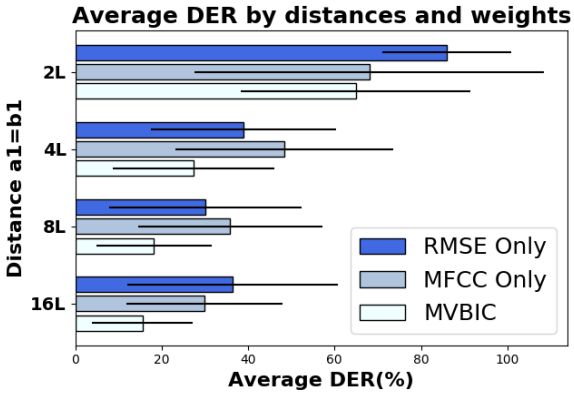

Fig. 2. Average DER by distances of interfering speakers from primary speakers.

where i is index of vector representations and M is the number of feature vector representations. If we consider all N_B BIC values between all N_S speech segments (N_B=$^{N_S}C_2$) from the session as given data points, we can calculate the sample variance of $\widehat{b}$ from given N_B BIC values from each of M features as below:

$$Var\left[\widehat{b}\right] = \mathbf{w}^T \Sigma_b \mathbf{w} \tag{4}$$

Here, we make an assumption that the noise random variable n_i is mean zero and the two noise streams are uncorrelated. This assumption mostly holds if the features are exploiting diverse information as is in the case of MFCC and RMSE. In addition, we also assume that the random variable b, which is the hidden and correct BIC value, and noise random variable n_i are uncorrelated. Thus, the M by M covariance matrix Σ_b in equation (4) has elements described as:

$$\begin{aligned} \sigma^2_{b,i} &= \sigma^2 + \sigma^2_{n,i} \\ \sigma_{b,ij} &= \sigma_{b,ji} = \sigma^2 \\ \text{where } i \neq j &\text{ and } i, j \in [1, M] \end{aligned} \tag{5}$$

where $\sigma^2_{b,i}$, σ^2 and $\sigma^2_{n,i}$ are variances of b_i, b and n_i respectively. Using the above assumptions and constraining the sum of weights to 1, we can rewrite the variance of $\widehat{b}$:

$$Var[\widehat{b}] = \left(\sum_{i=1}^{M} \omega_i\right)^2 \sigma^2 + \sum_{i=1}^{M} \omega_i^2 \, \sigma^2_{n,i} \tag{6}$$

$$= \sigma^2 + \sum_{i=1}^{M} \omega_i^2 \, \sigma^2_{n,i} \tag{7}$$

Thus, minimizing variance of $\widehat{b}$, we can also minimize the variance of noise $\sigma^2_{n,i}$ on the assumption we make while keeping the σ^2 intact. Thus, we can set up a minimization problem as:

$$\begin{aligned} \textbf{Minimize:} \quad & Var\left[\widehat{b}\right] = \mathbf{w}^T \Sigma_b \mathbf{w} \\ \textbf{Subject to:} \quad & \mathbf{w}^T \mathbf{1} = 1 \end{aligned} \tag{8}$$

The solution to the equation (8) would be given as below:

$$\widehat{\mathbf{w}} = \frac{\Sigma_b^{-1} \mathbf{1}}{\mathbf{1}^T \Sigma_b^{-1} \mathbf{1}} \tag{9}$$

With solution in equation (9), we estimate the weight in equation (3) to obtain $\widehat{b}$.

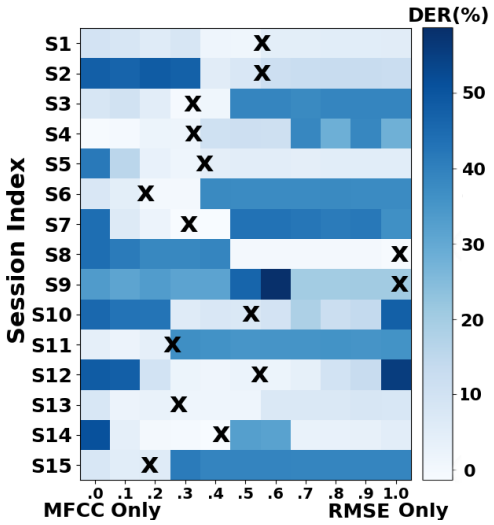

Fig. 3. The estimated weights ($\times$) layered on the results by each weight for the first 15 sessions.

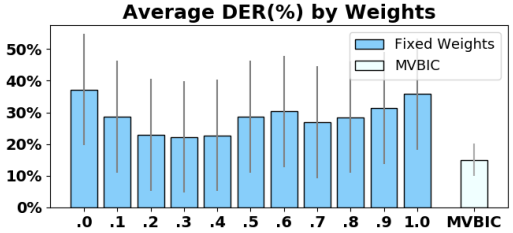

Fig. 4. Average DER by fixed weights and esimated weights from MVBIC for generated dataset.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

We verify the performance of the MVBIC method on USCDiarLibri2,4 and on RT-06S data. For the individual diarization streams we performed BIC value based clustering down to 4 clusters. All the experimental results below were tested with *md-eval* software in RT06S dataset [2]. The following results are only evaluated for the primary speakers.

4.1. The effect of the Distances of Interfering Speakers

In this experiment, we wanted to evaluate the effect of the distance of the interfering speakers from the microphone locations. For this experiment, we used a rectangular arrangement for the 4 speakers and generated 20 sessions per distance. We kept the distance between the two primary speakers fixed (to 5L) and varied the distance a_1 and b_1, as in Fig. 1, keeping $a_1 = b_1$. As Fig.2 shows, MVBIC keeps the DER lower than the single feature diarization methods regardless of the location of the interfering speakers. Furthermore, this experiment indicates that both features perform worse when interfering speakers are near the primary speakers. Importantly, we note that the distance of the interfering speaker greatly influences the relative accuracy of each of diarization stream and hence the weight of the stream should hold in case of fusion. This points further to the need for a dynamic fusion stream, such as MVBIC, as proposed above.

4.2. MVBIC evaluation on USCDiarLibri2,4

To verify the performance of proposed MVBIC technique, we randomly assigned the distance between all sources to be between 2 and 20 times L, as in Fig.1 and generated 50 sessions. Using this test dataset, the performance of proposed MVBIC method is compared with fixed BIC weights. In Fig. 3, the x-axis represents w, which is

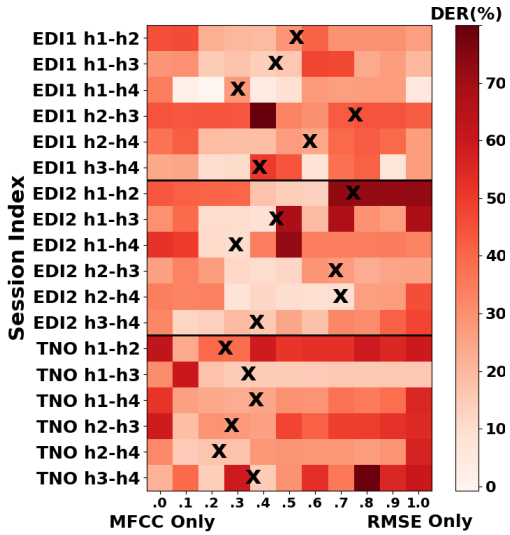

Fig. 5. The estimated weights ($\times$) layered on the results by each weight for the subset of RT06S dataset. Indexes h1-h4 refer to index of microphones.

the weight of the RMSE stream as follows: $\mathbf{w}^{\mathrm{T}} = [w_{\mathrm{RMSE}}, w_{\mathrm{MFCC}}] = [w, 1-w]$. We use $w = [0, 0.1, \ldots, 1.0]$. The DER results are visualized for each session and each weight. Note that to keep figure readable only the first 15 sessions are shown.

The "$\times$" marks in Fig. 3 describe BIC weights that MVBIC technique estimated for each session. We can see that the choice of $\mathbf{w}$ can play a significant role in the DER for each session. We also observe that the estimated weights by proposed MVBIC, marked "$\times$", are mostly tracking the minima of DER (whitest regions of each row). This outcome indicates that MVBIC can estimate, optimal according to our optimization criterion, values of the fusion vector from given BIC streams that result in near optimum fusion DER.

In Fig. 4, the DER results are shown for 50 sessions. The DER averages are plotted for the 50 for the different values of $\mathbf{w}$ as above. The last bar shows the result with the average DER based on the proposed MVBIC method. By optimizing a fixed $\mathbf{w}$ on the test set, we can see significant benefits over individual streams ($w = 0$ or $w = 1$) or equal weights ($w = 0.5$). The best performing value in this case would be $\mathbf{w}^{\mathrm{T}}$ =[0.3, 0.7]. However such optimization is not possible as the test data are not available at training time, but only serves as an upper bound for the static $\mathbf{w}$ fusion. The MVBIC method in contrast, even without optimization on the test data, it can beat any static fusion weight $\mathbf{w}$ as we can see from the last bar. This result shows that if the data is of high variability or mismatched to the training and development data, the proposed MVBIC can perform significantly better than a static, pre-tuned weight. Thus, the proposed technique can be an effective way to cope with the heterogenous data we observe in real-world conditions.

4.3. Evaluating on RT06S

We tested the performance of the proposed MVBIC system with individual head microphones for each session in RT06S dataset [2]. We picked three meetings (EDI1: EDI_20050216-1051, EDI2: EDI_20050218-0900, TNO: TNO_20041103-1130) which have the same number of total speakers in USCDiarLibri2,4. Among the four speakers in each meeting, two speakers are regarded as primary speakers and the rest of two speakers are regarded as interfering speakers. Thus, total 6 (4C2) microphone combinations were tested for each of the three meetings.

Fig. 5 shows the same type of visualization as Fig. 3. We see that

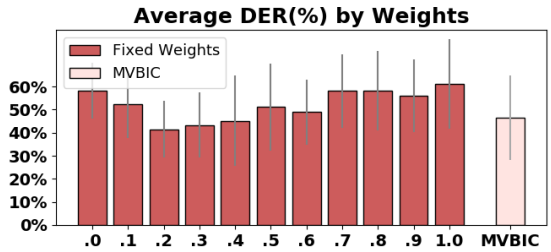

Fig. 6. Average DER by grid-searched weights and esimated weights from MVBIC for subset of RT06S dataset.

the MVBIC method does not pick as good candidates as we would expect. We also see that there seem to be multiple minima in the DER vs $\mathbf{w}$ space. This is likely due to the longer length of the sessions and the varying acoustic conditions. Since we only find one $\mathbf{w}$ using MVBIC per session, this is suboptimal.

Fig. 6 shows the result for RT06S dataset in the same format as 4. The proposed method showed 46.5% of DER while the most accurate fixed weight result showed 41.5% of DER. Again we observe that the MVBIC method approaches the optimize-on-test-set performance of the static weight.

Despite the highly mismatched conditions of this experiment, i.e. assuming stationary environment throughout the length of the session, which is false, and obtaining a single MVBIC weight $\mathbf{w}$ per session, and the higher-quality head-worn microphones, we still see significant benefits in using MVBIC.

5. CONCLUSIONS

We introduced a new dataset USCDiarLibri for evaluating Diarization algorithms that enables tunable task difficulty and conditions. We described and employed a subset of our proposed dataset. We also proposed a MVBIC method to estimate the fusion weights among multiple diarization streams. The proposed technique does not require any training data to determine the weights while it closely estimates the ideal weights, optimally according to the minimum variance criterion. This has significant benefits in real-world environments where the recording conditions are highly variable and heterogeneous. The proposed method allows also to exploit any available diarization stream dynamically, *i.e.,* increasing the fusion information streams if appropriate. In this work, we employed two diarization feature streams, RMSE and MFCC. A range of other information streams will be considered in future work such as multiple MFCC streams from each microphone, TDOA information, and lexical content similarly to our work in [6].

Further, any contributions by MVBIC are orthogonal to improvements in the individual diarization schemes and so newer methods, such as Deep Neural Network (DNN) based [26] or i-vector based methods, can be employed. The contributions are also generalizable to more sources, microphones, interferences *etc.* and will be evaluated further with the various conditions made possible by USCDiarLibri.

6. ACKNOWLEDGMENT

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 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] Jonathan G Fiscus, Nicolas Radde, John S Garofolo, Audrey Le, Jerome Ajot, and Christophe Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 369–389.

[2] Jonathan G Fiscus, Jerome Ajot, Martial Michel, and John S Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[3] Panayiotis G. Georgiou, Matthew P. Black, and Shrikanth S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: Some recent developments," in *Third International Workshop on Social Signal Processing (SSPW'11), ACM Multimedia'11*, Scottsdale, AZ, 2011, pp. 7–12.

[4] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1 –31, 2013.

[5] Bo Xiao, Panayiotis Georgiou, Zac E. Imel, David Atkins, and S. Narayanan, ""Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PLOS ONE*, December 2015.

[6] Bo Xiao, Chewei Huang, Zac E. Imel, David C. Atkins, Panayiotis Georgiou, and Shrikanth S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling," *PeerJ Computer Science*, vol. 2, pp. e59, Apr. 2016.

[7] Haoqi Li, Brian Baucom, and Panayiotis Georgiou, "Unsupervised latent behavior manifold learning from acoustic features: audio2behavior," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, March 2017.

[8] Shao-Yen Tseng, Brian Baucom, and Panayiotis Georgiou, "Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings," in *Proceedings of Interspeech*, Stockholm, Sweden, August 2017.

[9] Md Nasir, B. R. Baucom, Craig J Bryan, Shrikanth Narayanan, and Panayiotis Georgiou, "Complexity in speech and its relation to emotional bond in therapist-patient interactions during suicide risk assessment interviews," in *Interspeech*, Stockholm, Sweden, August 2017.

[10] M. Reblin, R. E. Heyman, L. Ellington, B. R. W. Baucom, P. G. Georgiou, and S. T. Vadaparampil, "Everyday couples' communication research: Overcoming methodological barriers with technology," *Patient Education & Counseling*, 2017 (in press).

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[12] Neville Ryant and Mark Liberman, "Automatic analysis of phonetic speech style dimensions.," in *Interspeech*, 2016, pp. 77–81.

[13] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulations and array processing algorithms," *arXiv preprint arXiv:1710.04196*, 2017.

[14] Tomas Gansler, Maria Hansson, C-J Ivarsson, and Göran Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communications*, vol. 44, no. 11, pp. 1421–1427, 1996.

[15] Jose M Pardo, Xavier Anguera, and Chuck Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Ninth International Conference on Spoken Language Processing*, 2006.

[16] Jose Pardo, Xavier Anguera, and Chuck Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, 2007.

[17] Jan Scheuing and Bin Yang, "Disambiguation of tdoa estimation for multiple sources in reverberant environments," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 8, pp. 1479–1489, 2008.

[18] Sue E Tranter and Douglas A Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[19] Sam Dutton, "Getting started with webrtc," *HTML5 Rocks*, vol. 23, 2012.

[20] Matthew A Siegler, Uday Jain, Bhiksha Raj, and Richard M Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, 1997, vol. 1997.

[21] Scott Chen and Ponani Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. darpa broadcast news transcription and understanding workshop*. Virginia, USA, 1998, vol. 8, pp. 127–132.

[22] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[23] James Lyons, "Python speech features," `https://github.com/jameslyons/python_speech_features`, 2017, Accessed: 2017-10-27.

[24] Bernd Scherer, "A note on the returns from minimum variance investing," *Journal of Empirical Finance*, vol. 18, no. 4, pp. 652–660, 2011.

[25] Robert G Lorenz and Stephen P Boyd, "Robust minimum variance beamforming," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1684–1696, 2005.

[26] Arindam Jati and Panayiotis Georgiou, "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.