ENHANCEMENT AND ANALYSIS OF CONVERSATIONAL SPEECH: JSALT 2017

Neville Ryant[a], Elika Bergelson[b], Kenneth Church[c], Alejandrina Cristia[d], Jun Du[e], Sriram Ganapathy[f], Sanjeev Khudanpur[g], Diana Kowalski[h], Mahesh Krishnamoorthy[i], Rajat Kulshreshta[j], Mark Liberman[a], Yu-Ding Lu[k], Matthew Maciejewski[g], Florian Metze[j], Jan Profant[l], Lei Sun[e], Yu Tsao[k], Zhou Yu[m]*

[a] Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Psychology and Neuroscience, Duke University, Durham, NC, USA
[c] IBM, Yorktown Heights, NY, USA
[d] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, Paris, France
[e] University of Science and Technology of China, Hefei, China
[f] Electrical Engineering Department, Indian Institute of Science, Bangalore, India
[g] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
[h] University of Illinois at Urbana-Champaign, Champaign, IL, USA
[i] Apple, Cupertino, CA, USA
[j] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[k] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[l] Brno University of Technology, Brno, Czech Republic
[m] Department of Computer Science, University of California Davis, Davis, CA, USA
* Corresponding author: nryant@ldc.upenn.edu

ABSTRACT

Automatic speech recognition is more and more widely and effectively used. Nevertheless, in some automatic speech analysis tasks the state of the art is surprisingly poor. One of these is “diarization", the task of determining who spoke when. Diarization is key to processing meeting audio and clinical interviews, extended recordings such as police body cam or child language acquisition data, and any other speech data involving multiple speakers whose voices are not cleanly separated into individual channels. Overlapping speech, environmental noise and suboptimal recording techniques make the problem harder. During the JSALT Summer Workshop at CMU in 2017, an international team of researchers worked on several aspects of this problem, including calibration of the state of the art, detection of overlaps, enhancement of noisy recordings, and classification of shorter speech segments. This paper sketches the workshop’s results, and announces plans for a “Diarization Challenge" to encourage further progress.

Index Terms— diarization, overlap detection, speech enhancement, automatic speech recognition

The research reported here was conducted at the 2017 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Carnegie Mellon University and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Apple, Facebook, Google, and Microsoft. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

1. INTRODUCTION

Digital audio is increasingly pervasive, and automatic speech recognition is more and more widely and effectively used, for interactive applications as well as for data mining and indexing of audio archives. Nevertheless, the state of the art for some automatic speech analysis tasks is surprisingly poor. One of these is “diarization", the task of determining who spoke when.

Diarization is key to processing meeting audio and clinical interviews, extended recordings such as police body cam or child language acquisition data, and any other speech data involving multiple speakers whose voices are not cleanly separated into individual channels, including both audio archives and multi-speaker interactive applications. Overlapping speech, environmental noise and suboptimal recording techniques make the problem harder. During the JSALT Summer Workshop at CMU in 2017[1], an international team of researchers worked on several aspects of this problem, including calibration of the state of the art, detection of overlaps, enhancement of noisy recordings, and the classification of shorter speech segments.

In this paper we sketch the workshop’s results. We begin by calibrating the performance of existing approaches and variant forms of those approaches (Section 3). An important background issue is the choice of scoring techniques and scoring parameters – the methods most commonly used in past diarization projects can give a overoptimistic picture of system performance, since large unscored “collars" at the edges of speech segments cause these methods to ignore short speech segments that are critical in some applications. In addition, the decision to ignore overlapped speech, or to score any of the simultaneous speakers as correct, can be a problem in highly inter-

[1] https://www.lti.cs.cmu.edu/2017-jelinek-workshop

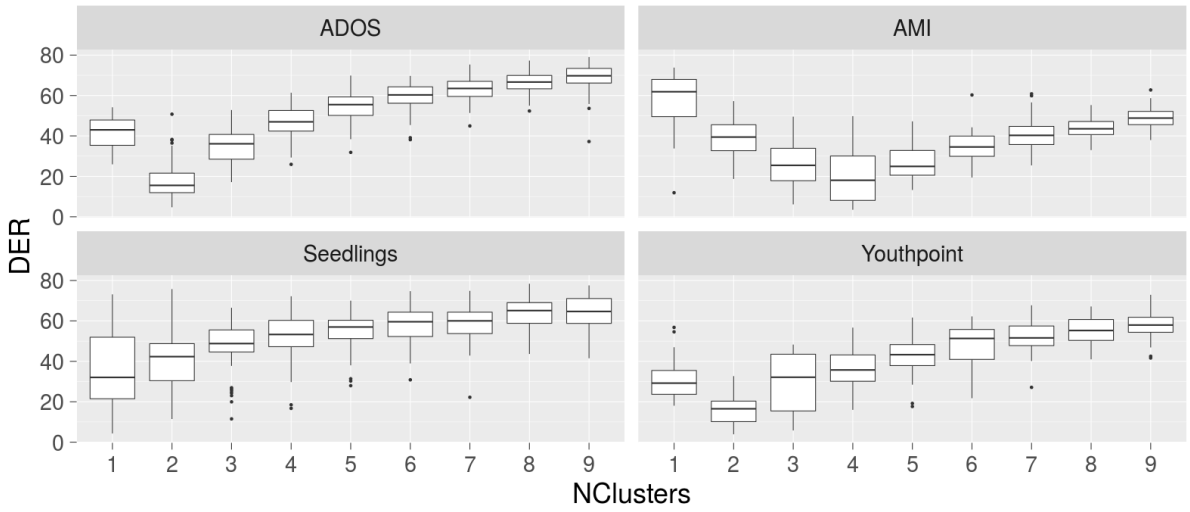

Fig. 1. Box-and-whiskers plots of recording level DER as a function of number of target clusters across four corpora for the VBDiarization system.

active conversations. We discuss how to avoid these problems, and present open-source code implementing several alternative scoring techniques and scoring parameterizations (Section 2).

Workshop researchers also explored several novel approaches to improving diarization. One of these projects explored the value in diarization tasks of speech enhancement via deep recurrent autoencoders (Section 4). Another project worked on improving the state-of-the-art in overlap detection (Section 5). Still another explored improvements to multi-channel far-field acoustic modeling using 3-D convolutional networks (Section 6). We also explored the idea that a functional analysis of (stochastic) conversational dynamics might help in performing the diarization task, as well as providing information of value in its own right. And finally, we explored the potential value of human-in-the-loop architectures, where a relatively small amount of human judgment can be used to improve performance by seeding a system with relevant training segments, constraining the number of speakers, or evaluating system-proposed segment clusters.

A wide variety of datasets were used in these explorations, including the AMI meeting corpus [1], a soon-to-be-published collection of Autism Diagnostic Observation Schedule (ADOS) interviews [2], henceforth referred to as ADOS, three collections of day-long child language acquisition recordings [3, 4, 5], a soon-to-be-published collection of broadcast interviews (YouthPoint), and a sample of neuropsychological testing interviews from the Framingham Heart Study (FHS) [6].

Progress in recent years on corpora such as CALLHOME[7], where diarization error rates (DER) have fallen below 10% [8], have lead some researchers to view diarization as a problem that, if not solved, is mostly solved. And for corpora such as CALLHOME, which consist of relatively clean conversational telephone speech, performance is indeed quite good. However, performance degrades markedly for other domains, as is clearly illustrated by Fig. 1, which depicts DER of one freely available, competitive i-vector based diarization system, VBDiarization[2], as a function of the target number of clusters on four corpora: the single distance microphone (SDM) condition of AMI, YouthPoint, Seedlings, and ADOS. For the latter two, which include abundant child speech, diarization is abysmal and even for YouthPoint, which consists of wideband, cleanly recorded interview speech and should be relatively easy, DER is poor for many recordings.

Given the success of diarization on CALLHOME and the relatively low error rates in recent Speaker Recognition Evaluations (SRE), this might surprise some researchers. So, as a test, we treated diarization as a speaker verification problem, by taking all utterances from the (gold standard transcription of the) AMI Dev meetings that are at least 3 seconds long, and using a state-of-the-art i-vector-based system to evaluate whether the two members of each pair came from the same speaker or not. Equal error rates ranged from 17.5% to 26.3%, depending on which microphones were used. For shorter segments, the performance is expected to be lower, because i-vector techniques become increasingly unstable as shorter windows are analyzed. More than 78% of the speech segments in the ADOS interviews are shorter than 3 seconds, and almost 30% are shorter than 1 second. In the FHS interviews, the proportion of short speech segments is even higher. And as the speech segments to be classified get shorter, we predict that the performance of i-vector based SRE techniques will fall nearly to chance level.

2. SCORING METRICS

The evaluation metric most commonly used for this task is “Diarization Error Rate" (DER) [9], which implements the simple and intuitive idea of measuring the fraction of analysis frames that are not correctly attributed, whether to a speaker or to non-speech. We have used this metric in reporting many of our results, given its status as the standard. However, there are several problems with this metric. First, the metric involves an unscored “collar" around the edges of speech segments, by default 250 ms, to allow inexactness of segment boundaries. This means that segments shorter than 500 ms are not scored at all, and longer segments are only partially scored. Second, the metric ignores overlapped speech, which may constitute a significant percentage of highly-interactive conversation (e.g. 30-50% of speech segments and about 12% of audio frames in the NIST

[2]https://github.com/Jamiroquai88/VBDiarization

	DER	tauYX	B3Prec	B3Rec	B3F1	H(X|Y)
DER	1.000	0.784	0.752	0.168	0.665	0.667
tauYX	0.784	1.000	0.995	0.459	0.959	0.973
B3Prec	0.752	0.995	1.000	0.469	0.996	9.988
B3Rec	0.168	0.459	0.469	1.000	0.677	0.514
B3F1	0.665	0.959	0.966	0.677	1.000	0.970
H(X|Y)	0.667	0.973	0.988	0.514	0.970	1.000

Table 1. Absolute value of correlation matrix for metrics in 100 ADOS Interviews (X=reference labeling and Y=system labeling).

meeting corpus [10]. And third, as Fig. 1 illustrates, the metric's results are highly unstable as a function of the number of hypothesized speakers, which itself is hard to get right. For example, a system that splits the frames belonging to a particular speaker into two equal but pure sets will get the same score as a system that assigns half of that speaker's frames to other random speakers or to non-speech.

Therefore we explored three other metrics. The first one is Goodman & Kruskal's tau [11], which estimates the fraction of variability in the categorical variable A that can be explained by the corresponding values of the categorical variable B. In our case this is the fraction of variability in the sequence of reference ("gold") speaker labels that can be explained by the sequence of hypothesized ("system") speaker labels. The second new metric is B-cubed [12] – given analysis frame F_i in hypothesized speaker category C with true speaker S, we define that frame's B-cubed precision as the proportion of all frames in C that correspond to S, and that frame's B-cubed recall as the proportion of all frames from speaker S that are in category C. Overall B-cubed Precision and Recall are then the mean precision and recall of all frames, and the overall B-cubed F-measure is the usual harmonic mean of precision and recall. Our third new diarization metric is simply Conditional Entropy, where H(X|Y) is the entropy of the discrete random variable X given the discrete random variable Y. In this case, X is the sequence of true frame-wise speaker labels, and Y is the sequence of hypothesized speaker labels.

As Table 1 shows, there is a high correlation between the tau, B-cubed F1, and Conditional Entropy measures. We've implemented all four of these diarization metrics in an open-source suite of Python tools available on GitHub[3].

3. IMPROVEMENTS IN I-VECTOR BASED DIARIZATION

We explored improvements to a state-of-the-art i-vector based diarization system similar to that of [8], which uses overlapping 1.5 second sliding-window i-vectors reduced with a conversation-dependent PCA and scored via probabilitistic linear discriminant analysis (PLDA) [13]. The i-vector extractor and PLDA were trained on 80 hours of the SDM condition of the AMI meeting corpus using a 3 second window size to encourage more stable i-vectors. Segmentation was performed using agglomerative hierarchical clustering (AHC) using an adaptive stopping criterion [13].

Since diarization is highly sensitive to the input speech segmentation, we chose to omit speech activity detection and instead use the oracle segmentation throughout the workshop. For similar reasons, we also omitted a final resegmentation stage. On the AMI SDM condition, the baseline system achieves a DER of 18.77%.

Long windows were found to improve the quality of the PLDA, despite this resulting in a mismatch between the window sizes of i-vectors seen in training and testing. During one experiment, we varied the extraction window from 10 to 180 seconds by combining speech segments in the training data. Performance was found to be highest with PLDA windows of 180 seconds: 17.68% DER.

We also examined the impact of training set size, by retraining the i-vector extractor on a large corpus of non-AMI wideband speech comprising SRE08, parts of Mixer 6 [14], and VoxCeleb [15]. Despite the resulting mismatched train/test conditions this caused, training the i-vector extractor on the external set exclusively reduced DER on AMI to 14.29%, a relative reduction in error of 23.87% compared to the baseline.

Recently, DNN embeddings have been proposed as an alternative to i-vectors for speaker representation [8], especially for shorter segments. Given that AMI contains a large proportion of short segments, which are known to be problematic for i-vectors, we also examined the impact of switching to DNN embeddings. Alone, DNN embeddings do about as well as more traditional i-vectors, 14.37% as compared to 14.29%, when trained on the same external corpus. However, the DNN embeddings appear to learn structure that is complementary to that learned by i-vectors. A fusion of the two representations yield large reductions in error compared to either representation alone. In a system using both i-vector and DNN features, DER on the AMI SDM condition is reduced to 9.84%, a relative reduction in error of 48.4% compared to the baseline and 31.14% compared to either representation alone.

4. SPEECH ENHANCEMENT

Traditional approaches to speech enhancement such as short term spectral amplitude (STSA) estimation [16] or a priori signal-to-noise (PSNR) estimation [17] are unable to effectively deal with non-stationary noise sources, rendering them insufficient for many real world speech environments. Moreover, the resulting speech often suffers from "musical noise" artifacts [18], which can actually reduce system performance. For instance, on the AMI corpus in the SDM condition, when using DiarTK [19] and the oracle number of speakers, DER rises from 29.73% 30.55% with STSA and to 36.06% with PSNR!

Recent work [20, 21, 22] has sought to overcome these deficits using supervised deep learning methods. See, for instance, [22], who train a long short-term memory (LSTM) [23] based autoencoder to reconstruct clean log power spectra from noisy spectra. When applied prior to diarization, these deep speech enhancement techniques reliably lead to lower DER. Indeed, on the SDM condition of AMI, the LSTM-DM model of [22] achieves a relative reduction in DER of 2.42%, despite the extreme mismatch between training and test conditions (trained on nearfield read newswire speech/tested on farfield meeting speech). ADOS represents a yet more extreme train/test mismatch given the presence of abundant child speech, totally absent during training, yet LSTM enhancement is able to realize a relative reduction of 10.15% in DER even for this corpus.

During the workshop we further improved LSTM-based enhancement with three modifications to the LSTM-DM model. First, following [21] intermediate targets are introduced into the network so that it must achieve progressively larger SNR gains at higher levels; that is, if the noisy input is 0 dB, the network has intermediate targets at 10 dB and 20 dB in addition to clean speech. Second, we utilize multitarget learning (MTL) [20], forcing the network to both reconstruct denoised signals AND predict the ideal ratio mask (IRM) [24]. Third, we borrow from computer vision by utilizing a DenseNet [25] inspired structure which connects each layer to every

[3]https://github.com/nryant/dscore

other layer in a feed-forward fashion . The resultant model achieves relative reductions in DER of 5.9% for AMI and 11.46% for ADOS. Compared to the simpler LSTM-DM model, these represent reductions of 3.56% and 1.47% relative. The network is implemented in CNTK and will be released as an open source tool.

5. OVERLAP DETECTION

The presence of overlapping speech in audio recordings poses a great challenge for many speech applications such as automatic speech recognition (ASR) [26], speaker identification (SID) and speaker diarization. For instance, speaker diarization relies on agglomerative clustering of i-vector features extracted from speech segments, each of which is assumed to come from a single speaker. When speech is overlapped within a segment and this remains undetected, the extracted i-vector comes from a different distribution, which poses obvious problems for clustering. As anywhere between 30% and 50% of segments (and 10-13% of all frames) in conversational speech – for instance, CALLHOME, the NIST meeting corpora, and AMI – are overlapped, undetected and unhandled overlapped speech constitutes a major source of error.

While ideally a system would not only detect overlap but also perform source separation to tease out the individual speakers, source separation is challenging in practice and in many cases the mere detection and omission of overlapped regions is sufficient to achieve reduced error rates [27]. For instance, on AMI, we observe a 27.55% reduction in DER (relative) for the baseline from Section 3 when overlaps are removed prior to diarization. However, the state-of-the-art in this area remains well below acceptable levels for use in a processing pipeline; see, for instance, [28], who report 67 % precision and 34% recall on synthetic overlapped speech produced by combining TIMIT [29] utterances. During the workshop, we explored a constellation of models for overlap detection, first in the context of artificial overlaps produced using TIMIT utterances in a way similar to [28], then with meeting speech taken from the SDM condition of the AMI corpus. A variety of input features were considered, including Mel filterbanks, gammatone filterbanks, kurtosis of the power spectrum, and Wiener entropy. We also investigated a number of different acoustic model architectures, including diagonal covariance gaussian mixture models (GMMs), fully-connected neural networks, shallow convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and convolutional long short-term memory, fully-connected deep neural networks (CLDNN) [30].

The best performance for the synthetic TIMIT overlap data was achieved using Mel filterbank features and an architecture consisting of a unidirectional LSTM and three fully connected layers. On TIMIT, this model achieved frame-level accuracy of 73.7% for non-overlapped speech and 83.1% for overlapped speech. Performance on the AMI SDM data was not quite as good, but the model still achieved a frame-level accuracy of 77.0% for non-overlapped speech and 68.0% for overlapped speech. Further improvements were achieved by adding a 3-state HMM on top of the acoustic model and using Viterbi decoding, resulting in 87.9% accuracy for non-overlapped speech and 71% accuracy for overlapped speech. Most importantly, when utilized as part of a preprocessing stage for a variant of the diarization system described in Section 3, DER on AMI was reduced from 19.4% to 15.2%, a relative improvement of 21.7%. More details on overlap detection can be found in [31].

6. MULTI-CHANNEL FAR-FIELD ASR

Even with the recent advancements on ASR, the recognition of far-field multi-speaker conversational speech is quite challenging. The problem arises mainly due to the artifacts present in the signal caused by reverberation as well as the presence of multiple speakers. The conventional method is to enhance the signal using delay-sum beamforming followed a single channel ASR system. During the workshop, we proposed a three-dimensional CNN architecture for multi-channel far-field ASR. The proposed architecture consists of a front-end that utilizes the time-frequency-channel dimensions of the input spectrogram to derive representations that are fed to a unidirectional LSTM. The models are trained with regularized version of sequence discriminative lattice-free maximum mutual information (MMI) cost function. Experiments are performed on the AMI database using the first three channels from the MDM condition microphone array. The proposed methods show significant improvements over the baseline system that uses beamforming of the multi-channel audio along with a 2-D conventional CNN framework (absolute improvements of 1.1% over the best beamformed baseline system on AMI dataset). More details of various ASR experiments can be found in [32]. These experiments suggest that end-to-end modeling with the ASR cost function may be more optimal than the conventional two-stage design of a beamforming method followed by a ASR acoustic modeling stage.

7. CONCLUSION

During the JSALT 2017 summer workshop, we explored several new approaches to diarization, and made some improvements in standard methods. But as we expected, the general problem is by no means solved. So to encourage further progress, we have organized a new annual diarization challenge, DIHARD, the first of which is being held in conjunction with Interspeech 2018. The goals of this challenge are threefold: (1) to create an evaluation set drawn from a diverse set of challenging domains; (2) to establish a baseline of performance for existing diarization technologies on this set; (3) to release the reference data and results for continued research after the evaluation to encourage further testing and development.

DIHARD will focus on "hard" diarization in challenging corpora where the expectation is that current state-of-the-art will fare poorly. The materials will consist of short (5-10 minute) single-channel recordings involving various numbers of speakers from a wide variety of domains, including clinical interviews, child language recordings, business meetings, web video, conversations in restaurants, courtroom discussions, sociolinguistic interviews, and audiobooks. Two tracks are supported: (1) given the audio and a reference SAD, perform diarization; (2) given only the audio, perform diarization from scratch. Both tracks will be evaluated in terms of DER and frame-wise mutual information.

Future Challenges might include such things as the use of multiple audio channels, analysis of conversational dynamics, grounded diarization where (a smaller or larger amount of) training material is provided for some speakers, evaluation of human-in-the-loop efficiency for alternative methods, and so on.

8. REFERENCES

[1] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[2] J. Parish-Morris, M. Liberman, N. Ryant, C. Cieri, L. Bateman, E. Ferguson, and R. T. Schultz, "Exploring Autism Spectrum Disorders using HLT.," in *CLPsych*, 2016, pp. 74–84.

[3] M. Casillas, "Casillas-X-cultural (from The Language Archive," https://hdl.handle.net/1839/9E3EF620-690E-4BC1-8A10-B6815AF84DAB@view, Accessed: 2017-08-22.

[4] E. Bergelson, A. Warlaumont, A. Cristia, C. Rowland, M. Casillas, C. Rosemberg, M. Soderstrom, F. Metze, E. Dupoux, and O. Rasanen, *Starter-ACLEW*, Databrary, 2017.

[5] E. Bergelson, *Bergelson Seedlings HomeBank Corpus*, Accessed: 2017-08-22.

[6] T. R. Dawber, G. F. Meadors, and F. E. Moore Jr, "Epidemiological approaches to heart disease: the Framingham Study," *American Journal of Public Health and the Nations Health*, vol. 41, no. 3, pp. 279–286, 1951.

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

[8] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *ICASSP*, 2017, pp. 4930–4934.

[9] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 369–389.

[10] O. Cetin and E. Shriberg, "Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I357–I360.

[11] L. A. Goodman and W. H. Kruskal, "Measures of association for cross classifications," *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 732–764, 1954.

[12] A. Bagga and B. Baldwin, "Algorithms for scoring coreference chains," in *LREC*, 1998, vol. 1, pp. 563–566.

[13] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *SLT*, 2014, pp. 413–417.

[14] L. Brandschain, D. Graff, and K. Walker, *Mixer 6 Speech LDC2013S03*, Linguistic Data Consortium, Philadelphia, 2013.

[15] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 32, no. 6, pp. 1109–1121, 1984.

[17] P. Scalart et al., "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, 1996, vol. 2, pp. 629–632.

[18] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, pp. 873–902. Springer, 2008.

[19] D. Vijayasenan and F. Valente, "DiarTk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *INTERPSEECH 2012*, 2012.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[21] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement.," in *INTERSPEECH*, 2016, pp. 3713–3717.

[22] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *HSCMA*, 2017, pp. 136–140.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[25] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[26] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation.," in *INTERSPEECH*, 2001, pp. 1359–1362.

[27] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *ICASSP*, 2008, pp. 4353–4356.

[28] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. H. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *ICASSP*, 2013, pp. 2834–2838.

[29] J. Garofolo, L. Lamel, W. Fisher, D. Pallet, and N. Dahlgren, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, 1993.

[30] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *ICASSP*, 2015, pp. 4580–4584.

[31] N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging LSTM models for overlap detection in multi-party meetings," in *ICASSP*, 2018.

[32] S. Ganapathy and V. Peddinti, "3-d cnn models for far-field multi-channel speech recognition," in *submitted to ICASSP*, 2017.

[33] N. A. Nystrom, M. J. Levine, R. Z. Roskies, and J. Scott, "Bridges: a uniquely flexible HPC resource for new communities and data analytics," in *XSEDE*, 2015, p. 30.

[34] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, et al., "XSEDE: accelerating scientific discovery," *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, 2014.