MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

# Scene-aware Interaction

Hori, Chiori; Tsuchiya, Masato; Chen, Siheng; Cherian, Anoop; Hori, Takaaki; Harsham, Bret A.; Marks, Tim K.; Le Roux, Jonathan; Sullivan, Alan; Vetro, Anthony

TR2021-042 May 04, 2021

**Abstract**

The recent artificial intelligence (AI) boom and intelligent use of data acquired from various sensors has accelerated the development of technologies needed to realize advanced human-like capabilities in machines. AI technologies have come a long way in accurately perceiving visual scenes and understanding speech. However, one important piece of technology is still missing: natural and context-aware human-machine interaction, where machines understand their surrounding scene from the human perspective and are able to share their understanding with humans using natural language. To bridge this communication gap, we have developed and built a new AI system, called scene-aware interaction, that enables machines to translate their perception and understanding of a scene and respond to it using natural language to interact more effectively with humans. This paper introduces an example application of scene-aware interaction to car navigation systems, which will provide drivers with intuitive route guidance.

*Society of Automotive Engineers of Japan*



Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

# マルチモーダルセンシング情報に基づく Scene-aware Interaction 技術

堀 智織[2]・土屋政人[1]・Siheng Chen・Anoop Cherian[2]・堀 貴明・Bret Harsham[2]
Tim K. Marks[2]・Jonathan Le Roux[2]・Alan Sullivan[2]・Anthony Vetro[2]

[1]三菱電機情報技術総合研究所、[2]Mitsubishi Electric Research Laboratories

## １．はじめに

約 100 年前に「ロボット」という人間を模した機械が人間の代わりに労働するという物語が創作されて以来、人間と会話し、人間のように自律行動できるロボットが多くの映画やアニメなどに登場するようになった。ロボットの形状は人型に留まらず自動車型など多様で、ロボットは個性を持つ人間として擬人化されて描かれている。現実の世界では AI 技術の革新的な進歩と様々なセンサから取得されたデータの知的利用により、ロボットの実用化に向けた要素技術の開発が急速に加速している。物体認識・動画説明・自然言語生成・音声対話技術の性能は深層学習を用いることで劇的に改善し、タスクによってはすでに人間を超える能力を示している。例えば、多言語複数話者による同時発話の音声認識技術[1]などがその良い例である。このような技術革新の結果、コンピュータは人間の言葉を理解するようになり、人間の指示に従いアプリケーションを操作し、インターネット上にある画像や音声もキーワードで瞬時に検索できるようになった。しかし、実世界では SF 映画のロボットのように人間と同等の能力を持ち人間と換わることができる機械は未だ発明されていない。人間は実世界の動的に変化する複雑な事象を感覚器官から得られた情報を用いて理解し、さらにその理解を自然な言葉で表現して他の人間と意思疎通することができるが、現在の機械にはその重要で欠かすことのできない機能が無い。機械が実世界の事象を認識・理解し自然言語で意思疎通をするには、数値で表現された複数のセンシング情報を統合して一つの事象として理解し、それらが組み合わさった複雑な事象を人間の状況理解に対応した自然言語で表現し、さらに人間との対話の文脈と認識した周囲の状況を同時に考慮して適切な応答文を生成する必要がある。既存の技術では、個別の小規模な事象の認識に限定的な言語表現をクラス名として対応させる機能に留まり、自然言語生成を用いた詳細な状況理解には程遠かった。ところが、近年の深層学習の進歩によりニューラルネットワークでモデル化された識別器（クラス分類）や単語を並べて文を生成するような記号列生成装置、あるいはそれらを統合したモデルが単一のネットワークで表現できるようになり、End-to-End 学習を用いて入出力のサンプルだけで学習できるようになった。この技術的革新により、ロボットが必要とする実世界の状況理解に基づく音声対話機能が実現できる道筋が見え、我々は End-to-End 深層学習を用いて、複数のセンサが収集した情報（マルチモーダルセンシング情報）から周囲の状況を機械が自然な言語で理解し、人とより円滑な意思疎通を実現する「Scene-Aware Interaction 技術」を開発することができた。Scene-aware interaction 技術は、カメラで撮影した画像情報、マイクロフォンで集音した音響情報、LiDAR やレーダーで取得した位置情報などのマルチモーダルセンシング情報から、何がどこでどのような状態にあるのか、誰がどこで何をしているのか、といった周囲の状況を機械が自然言語で理解し、人間との会話の文脈も考慮して応答文を生成する技術である。本 Scene-aware Interaction 技術は、ロボットやモニタリングシステムといった状況理解に基づき人間とインタラクションを必要とする様々なシステムへの応用が期待できる画期的な技術である。本稿では、その応用例として車載の様々なセンサから取得された複数のセンサ情報に基づき状況を理解し経路案内を行うシステムを紹介する。

## ２. 未来型カーナビゲーションシステム

近年、多くのカーインフォテインメントシステムの操作に音声対話が導入されている。音声対話は運転操作を妨げることのない非常に有効なインタラクション手段であるが、現在システムが扱える知識はコンピュータ上に保存された既存の知識かネットなどから配信される交通状況など限定的である。市販のカーナビゲーションシステムの中には詳細化された地図情報に基づき商業施設名で案内するものもあるが、未だ GPS の自車位置情報に基づき「１００ｍ先の交差点を右に曲がってください。」といった距離を用いた経路案内するシステムが多い。一方、助手席の人間が経路案内をする場合、「郵便ポストのところで右折してください。」「交差点で左折している銀色の車に続いて左折してください。」のように今見えている特徴的な目印となる物体を用いて直感的に説明し、さらに進行方向に危険が迫れば「歩行者が道を横切ろうとしています。」といったように注意喚起もしてくれる。助手席の人間は、刻々と変化する状況に応じて運転者が認識しやすい物体を用いて経路案内をする。このような直感的な経路案内が可能になるのは、運転者と助手席の人間が同じ環境下で状況に対する認識と理解を共有しているからである。機械で助手席の人間と同等の機能を実現するには、実世界の動的に変化する事象を認識し、人間の視座で捉えた状況理解を行い、人間と自然言語で意思疎通をする Scene-aware Interaction 技術が必要となる。図１に Scene-aware Interaction 技術を用いた経路案内システムの概要を示す。

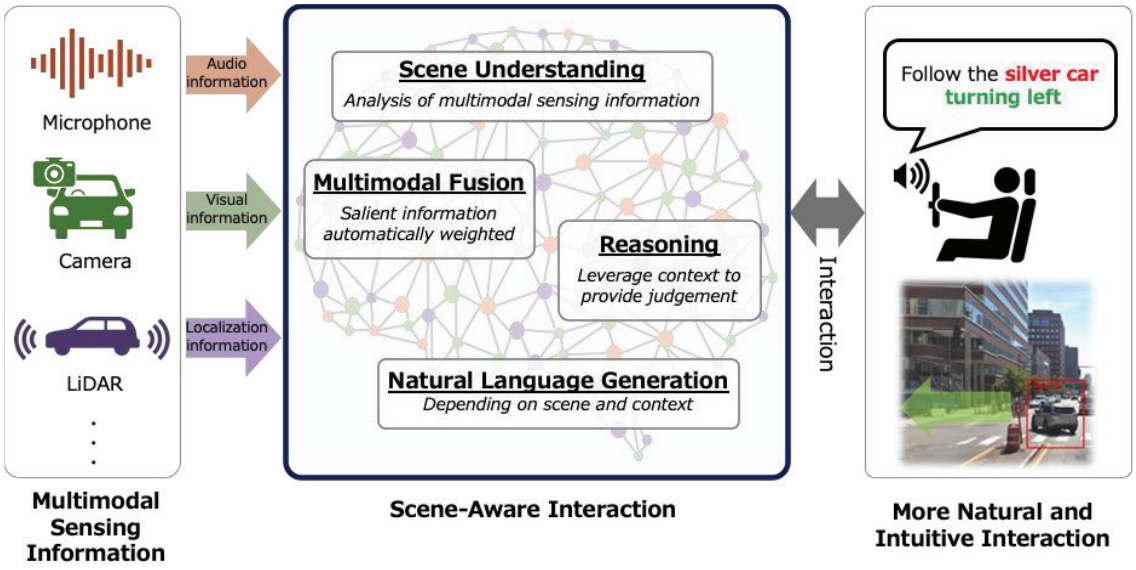


図１　Scene-aware Interaction 技術を用いた経路案内システムの概要

現在、大多数の自動車が安全のための運転支援と自動制御を目的として多くのセンサを搭載している。車外にカメラ、ミリ波レーダ、超音波センサが装備され、自動走行制御、自動緊急ブレーキ、車線維持及び駐車補助に用いられており、車内カメラは運転者の健康状態をモニタする。これらのセンサは死角にある車の存在をビープ音で、車線逸脱をハンドルの振動で運転者に知らせるが、実世界の状況について自然言語を用いて機械が運転者とインタラクションを行う目的では用いられていない。運転者とインタラクションするのにどのような方法が良いのかという議論があるが、最近の研究では自然言語で表現された詳細なメッセージがより効果的に意味情報を伝達するということが示されており、緊急事態を除けば、直感的で特別な訓練を必要としないことから最も望ましく、一般に受け入れられている方法であると報告されている[2]。

## ３．マルチモーダルセンシング情報に基づく状況理解

実世界の事象について自然言語でインタラクションを行うシステムにおいて最も重要な機能は、物体や事象を認識する機能と、そのセンシング情報を人間の理解に対応した自然言語に翻訳する機能である。本章では、画像認識や音響認識などの複数のセンシング情報を用いた物体・事象認識技術、及びそれらの認識内容を自然言語に変換する系列変換技術を紹介する。

### 3.1 画像特徴量

直感的な経路案内に欠かすことができないセンシング情報としてカメラの映像がある。助手席の人間による案内を模した未来型のカーナビゲーションシステムは、画像情報から進行方向を特徴づけるランドマークを探し、自車とランドマークあるいはランドマーク間の位置関係を用いて案内文を生成し、進路を遮る物体があれば注意喚起を促す文を生成する必要がある。これらのシステムの生成文で用いられる物体や事象の認識は、物体認識・動体トラッキング（Object Detection and Tracking）[3]、意味的領域分割(Semantic Segmentation)[4][5]、深度推定(Depth estimation)、Scene-graph 表現を用いた物体間の位置関係推定[6]、3 次元点群(3D Point Cloud)に基づく動作推定[7]などの画像処理技術によって実現可能である。図２に車載カメラの画像処理結果の例を示す。


図２　状況理解の根拠となる特徴抽出を支える画像処理技術

左上：車載カメラ映像の前方方向の風景における様々な種類の物体認識と追跡結果、右上：自動車、道路、歩行者といったピクセル単位での意味的領域分割結果、左下：自動車と他の物体の距離の関係を示す深度推定結果;右下：Scene-graph 表現に基づく物体間の位置関係の推定結果

本システムは、運転者が直感的に認識することができるランドマークを用いて経路案内を行うことから、システムは物体認識された膨大な数の物体の中から最も直感的に認識できる物体を重要な物体として選択する必要がある。ランドマークとしての重要度は、物体認識の結果に基づき物体の種類、属性、サイズ、位置、交差点からの距離、および進行方向などの特徴に基づきニューラル回帰モデルを用いて決定される。距離が遠ければ大きく目を引く物体、左折の際は交差点の左側にある他とは特徴が異なる物体、十分に特徴的なランドマークが無い場合には従来通り交差点からの距離を用いる。店の看板や自動車のロゴなどを認識することより、さらに詳細な経路案内が可能となる。さらに動体トラッキングの特徴と組み合わせることで、進行方向と同方向に進む前方方向の車を特定して目標物とし、進路に進入する物体を予測して優先的に注意喚起を行う。図３に動体トラッキングを用いた経路案内の例を示す。


図３. 動体トラッキングを用いた経路案内と注意喚起

ランドマークの複数の候補の中から最も特徴的な物体を選ぶ際、最も考慮すべき情報は物体間の距離に基づく位置関係である。現在の GPS や単眼カメラのみを用いた物体の位置関係推定には限界があり、将来的に GNSS 受信機や C-V2X といった道路インフラとの通信により高精度な自車位置情報の取得が期待できるが、現時点で利用することはできない。今回の実験では、自動運転技術のための 3D マッピング技術を用いて、カメラ画像と LiDAR 画像から 3D マップ上で物体の位置を推定する方法を適用することとした[8]。我々は LiDAR、カメラ、IMUs を搭載した 3D マップ作成のための Mobile Mapping System (MMS)を用いてデータ収集を行なった。図３に MMS で取得したデータに基づく物体の位置推定の例を示す。カメラ画像で認識された物体と LiDAR で取得した 3D ポイントクラウド上の同一の物体を対応づけることにより、3D ポイントクラウドから再構築されたトップダウンマップ上で物体の正確な位置を取得することが可能となる。

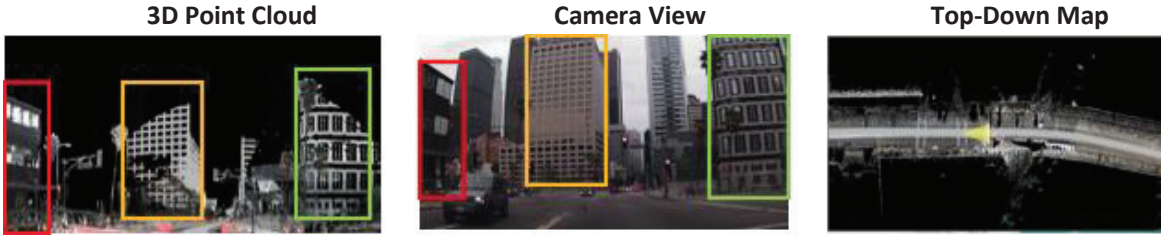


図３.MMS を用いて取得された３D ポイントクラウドと
カメラ画像およびトップダウンマップ

### 3.2. マルチモーダルセンセシング情報

人間と同等以上に周囲の事象を認識するためには、画像情報に留まらず、音響情報など様々なセンシング情報も考慮する必要がある。例えば、画像で救急車は認識できていないが音響情報からその動向が推定でき、「後方から救急車が近づいて来ます。」といった説明文が生成できる。この場合「後方」「救急車」という単語は画像の特徴量のみで生成することはできない。マルチモーダルセンシング情報を用いることで人間同等あるいはそれ以上の状況理解が可能になる。複数の信号をまとめ上げて一つの事象として理解するには、マルチモーダルセンシング情報を統合して新たな意味空間を構築し、その空間上で単語列を特徴づける必要がある。我々は単語毎に特徴的なセンシング情報に異なる重みがかかるように自動学習するマルチモーダル・アテンション法を考案した[11]。その結果、異なるセンシング情報が相互補完することで、より正しく状況理解することが可能となる。

### 3.3 系列変換モデル

複数のセンシング情報から単語列を生成する方法として、系列から系列に変換する系列変換モデル（Sequence-to-sequence model）が提案された。原言語の単語列を入力し目的言語の単語列を出力する機械翻訳[9]や音声信号系列を入力として書き起こしの単語列を出力する音声認識[10]などに盛んに用いられ、動画における事象を自然言語で説明する動画説明（Video Description）技術にも適用された[11]。これは、動画の特徴量を入力として説明文の単語列を出力することで実現したものである。我々は画像情報だけでなく音響情報もマルチモーダル・アテンション法を用いて考慮することにより、YouTube 動画説明の共通タスクで性能が改善されることを示した[12]。

## ４. ニューラル対話モデルに基づく Scene-aware Interaction

音声対話システムの研究は人手によるユーザ意図のラベル付けが必須のためシステムを大規模化することが非常に難しいという問題があった。近年の系列変換の深層学習モデルにより、対話の履歴とユーザの入力に基づき、システムの応答を自動生成する技術が検討されている。我々は 2016 年より Dialog System Technology Challenge (DSTC)を主催し、カスタマーサービスのテキストチャットなどの対話データを用いて End-to-End 深層学習による応答自動生成の研究開発を行なっている[13]。しかしながら、これらの対話システムは全ての情報がテキスト化された学習データであることを前提としており、逐次的に変化する周囲の状況をセンシングした情報に基づく対話を対象としていない。我々は、ニューラルネットワークに基づく動画説明技術とこの対話技術を一体化して End-to-End で学習させる Scene-Aware Interaction 技術を考案し、Audio-Visual Scene-aware Dialog (AVSD)という研究課題に取

り組んでいる。図４に AVSD のためのニューラルネットワークの構成を示す。本研究課題を発展させるため、DSTC 主催のワークショップにて AVSD をタスクとした競争的研究開発を行ない[14][15]、我々は図５に示す Dynamic Representation Learning を提案している[6]。

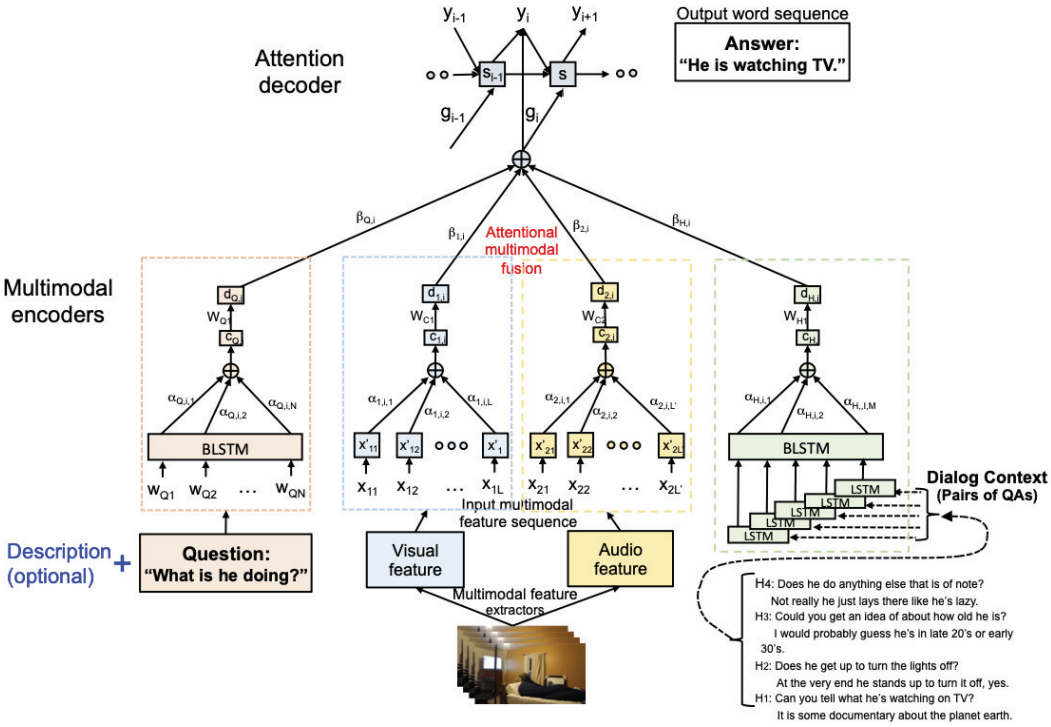


図４．AVSD のためのニューラルネットワークの構成

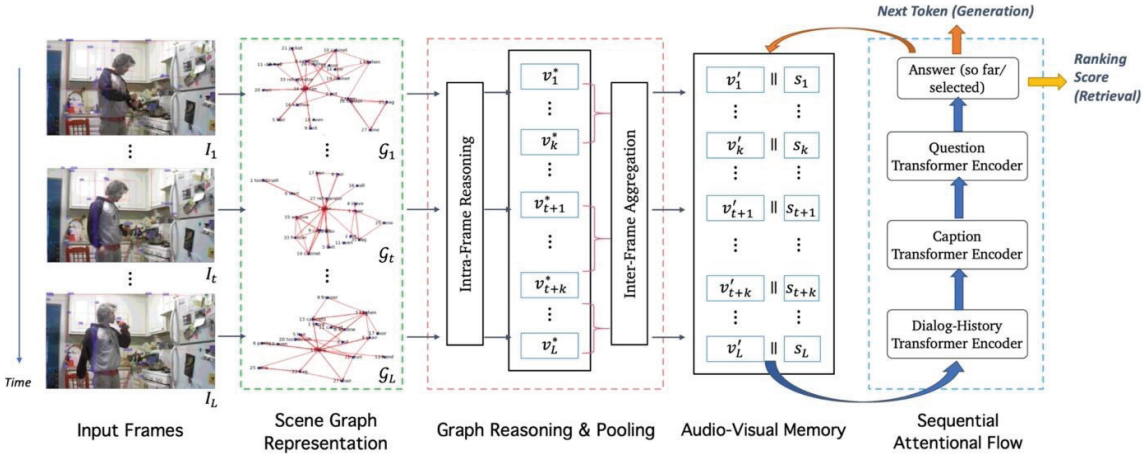


図５．AVSD のための Dynamic Representation Learning

## 5. 今後の展開

本稿では Scene-aware Interaction 技術を用いた次世代のカーナビゲーションシステムを紹介した(https://www.youtube.com/watch?v=zcA6p4DEIHU)。Scene-aware Interaction 技術は、自動運転への応用、無人施設の監視、公共空間における見守り、独居シニアのリモートサポートや、ロボットが人間と共同で作業をする場面など、様々な利用が期待されている。

## 参考文献


[1] Seki, H., Hori, T., Watanabe, S., Le Roux, J., Hershey, J.R., “End-to-end multilingual multi-speaker speech recognition,” in Proc. ISCA Interspeech, Sep. 2019, pp. 3755–3759.

[2] Bazilinskyy, P., Petermeijer, S.M., Petrovych, V., Dodou, D. and de Winter, J.C.,“Take-over requests in highly automated driving: A crowdsourcing survey on auditory, vibrotactile, and visual displays,” Transportation Research Part F: Traffic Psychology and Behaviour, Volume 56, 2018, pp. 82–98.

[3] Broad, A., Jones, M., & Lee, T. Y., “Recurrent Multi-frame Single Shot Detector for Video Object Detection,” In *BMVC* (p. 94), 2018, September.

[4] Yu, X., Chaturvedi, S., Feng, C., Taguchi, Y., Lee, T.-Y., Fernandes, C., Ramalingam, S., "VLASE: Vehicle Localization by Aggregating Semantic Edges," in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2018, pp. 3196-3203.

[5] Yu, Z., Feng, C., Liu, M.-Y., Ramalingam, S., "CASENet: Deep Category-Aware Semantic Edge Detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017.

[6] Geng, S., Gao, P., Chatterjee, M., Hori, C., Le Roux, J., Zhang, Y., Li, H., Cherian, A., "Dynamic Graph Representation Learning for Video Dialog via Multi-Modal Shuffled Transformers," in Proc. AAAI Conference on Artificial Intelligence, Feb. 2021.

[7] Wu, P., Chen, S., "MotionNet: Joint Perception and Motion Prediction for Autonomous Driving Based on Bird’s Eye View Maps," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020.

[8] Chen, S., Liu, B., Feng, C., Vallespi-Gonzalez, C., Wellington, C., "3D Point Cloud Processing and Learning for Autonomous Driving," IEEE Signal Processing Magazine, May 2020.

[9] Bahdanau, D., Cho, K., and Bengio, Y., "Neural machine translation by jointly learning to align and translate," in Proc. International Conference on Learning Representations (ICLR), May 2015.

[10] Chorowski, J., Bahdanau, D., Cho, K., and Bengio, Y., "End-to-end continuous speech recognition using attention-based recurrent NN: First results," in Proc. NIPS 2014 Workshop on Deep Learning, Dec. 2014.

[11] Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A., “Deep captioning with multimodal recurrent neural networks (m-RNN),” in Proc. International Conference on Learning Representations (ICLR), May 2015.

[12] Hori, C., Hori, T., Lee, T., Zhang, Z., Harsham, B., Hershey, J.R., Marks, T.K., Sumi, K., “Attention-Based Multimodal Fusion for Video Description,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 4193-4202.

[13] Hori, C., Perez, J., Yoshino, K., Kim, S. (Eds.), "Results from the 6th Dialog System Technology Challenge (DSTC6)" (Special Issue), Computer Speech and Language, 2019.

[14] Hori, C., Yoshino, K., D'Haro, L.F., Galley, M., Polymenakos, L.C. (Eds.), "The 7th Dialog System Technology Challenge (DSTC7)" (Special Issue), Computer Speech and Language, 2020.

[15] Kim, S., Schulz, H., Gunasekara, C., Hori, C., Rastogi, A., D'Haro, L.F. (Eds.), "The 8th Dialog System Technology Challenge (DSTC8)" (Special Issue), IEEE/ACM Transactions on Audio Speech and Language Processing, to appear in 2021.