# The next generation of audio intelligence: A survey-based perspective on improving audio analysis

BJÖRN SCHULLER[1,2,3], SHAHIN AMIRIPARIAN[2], GIL KEREN[2], ALICE BAIRD[2], MAXIMILIAN SCHMITT[2] AND NICHOLAS CUMMINS[2,*]

[1] *GLAM – Group on Language, Audio & Music, Imperial College London, SW7 2AZ London, UK*

[2] *Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, 86159 Augsburg, Germany*

[3] *audEERING GmbH, 82205 Gilching, Germany*

Computer audition has made major progress over the past decades; however it is still far from achieving human-level hearing abilities. Imagine, for example, the sounds associated with putting a water glass onto a table. As humans, we would be able to roughly "hear" the material of the glass, the table, and perhaps even how full the glass is. Current machine listening approaches, on the other hand, would mainly recognise the event of "glass put onto a table". In this context, this contribution aims to provide key insight into the already made remarkable advances in computer audition. It also identifies deficits in reaching human-like hearing abilities, such as in the given example. We summarise the state-of-the-art in traditional signal-processing-based audio pre-processing and feature representation, as well as automated learning such as by deep neural networks. This concerns, in particular, audio diarisation, source separation, understanding, but also ontologisation. Based on this, concluding avenues are given towards reaching the ambitious goal of "holistic human-parity" machine listening abilities – the next generation of audio intelligence.

## INTRODUCTION

Typical real-world audio consists of complex combinations of overlapping events from a variety of sources, creating both clashing and harmonious relationships. Despite this complexity, humans can, with relative ease, decipher across audio through understanding, decomposing, interpreting, and ontologisation an abundance of potentially conveyed messages and their related semantic meanings.

Historically, developments in the field of computational audio understanding (computer audition) were initially driven by speech analysis, in particular, the field of automatic speech recognition (ASR). From its inception at Bell labs in the 1950's with the "Audrey" system, capable of recognising spoken digits (Davis *et al.*, 1952), through the considerable advancements during the 1980's associated with the use of Hidden Markov

*Corresponding author: schuller@ieee.org

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019.  ISSN: 2596-5522.

Models (Hansen and Hasan, 2015), and to the recent deep learning revolution (Hinton *et al.*, 2012), ASR technologies have now matured to the point where they are embedded in everyday technologies, e. g., SIRI™, CORTANA™, and ALEXA™. A similar transforming effect has recently occurred through deep learning, in terms of the immense increase in recognition accuracy and robustness in music analysis (e. g., Coutinho *et al.*, 2014), and for the recognition of acoustic scenes and the detection of specific audio events (Mesaros *et al.*, 2018).

Considering the advances in computer audition throughout the last decades, the time is now to unite these domains of audio understanding by creating a fully-fledged across-audio approach, thereby pushing this somewhat overlooked and currently underdeveloped mode of research to the forefront of intelligent machine understanding. To date, computer audition approaches have been typically mono-domain focused, with only consideration for the previously aforementioned domains of speech, music, and in general in an isolated singular manner. The view proposed here would unify these domains to truly understand and interpret audio.

The ground-breaking nature of such an "across-audio analysis" approach is the simultaneous understanding of the entire acoustic scene. Imagine, as an example, an acoustic scene set in a garage with two people working on repairing a car while listening to music. An across-audio analysis will isolate the conversation, the music, and engine noises and then assign relevant state and trait tags to each. For instance, the music genre and individual instrumentation could be recognised, the age and gender of each person and their relationship to one another determined, the car's age, model and condition identified, and finally the repair duration logged.

In the following, we move quickly through the state-of-the-art in audio analysis as related to the needed aspects of such a view on the next generation of audio intelligence: audio diarisation, (audio) source separation, audio understanding, (audio) ontologisation.

## STATE-OF-THE-ART IN AUDIO ANALYSIS

### Audio diarisation

Audio diarisation is a generalisation of speaker diarisation to general sound sources, e. g., vehicles, musical instruments, animals, or background noise types (Reynolds and Torres-Carrasquillo, 2005). The state-of-the-art is mostly marked by speaker diarisation, as general audio diarisation is only gaining momentum at this time. Speaker diarisation thereby is tagging an audio recording of several individuals with speaker turn information, i. e., to provide information relating to "who is speaking when". The dominating trend of the last few years in speaker diarisation research is to find suitable speaker embeddings which give a reliable multi-dimensional clustering of speech segments according to speakers. In this regard, the *i-vector* and *Gaussian mixture model-based* approaches (Anguera *et al.*, 2012) are being overtaken by deep neural network (DNN) feature representations (Bredin, 2017). Note that DNN-based

speaker embeddings are sometimes called *d-vectors*, as opposed to *i-vectors* (Wang *et al.*, 2017). The advantage of DNNs for speaker diarisation is that they are capable of simultaneously learning the embeddings, i. e., the feature vectors describing speaker characteristics, and the scoring function, which represents the similarity between the embeddings of different segments (Garcia-Romero *et al.*, 2017). Nevertheless, when comparing different scoring functions for i-vector embeddings, DNNs have been shown to outperform conventional scoring functions, such as *cosine similarity* and *probabilistic linear discriminant analysis* (Le Lan *et al.*, 2017).

**Source separation**

Audio source separation is the decomposition of an arbitrary audio signal into several signals with only a single audio source of interest present in each decomposed part. The audio source could be a speaker, a musical instrument, a sound produced by an animal or a vehicle, or background noise, such as breaking sea waves. In most conventional approaches, a mixture-spectrogram is separated into several source spectrograms. In the past, nonnegative matrix factorization (*NMF*; Nikunen *et al.*, 2018) or *non-negative tensor factorisation* (Ozerov *et al.*, 2011) have been used for single-channel (monaural) source separation (Virtanen, 2007), and *independent component analysis (ICA)* or *multichannel NMF* (Nikunen *et al.*, 2018) used for multi-channel audio.

Well-studied aspects of source separation are speech denoising and speech enhancement. Previous research on speech denoising comprises *NMF* (Weninger *et al.*, 2012), *deep NMF* (Le Roux *et al.*, 2015), recurrent neural network (*RNN*)-based discriminative training (Weninger *et al.*, 2014b), *long short-term memory recurrent neural networks* (*LSTM-RNNs*; Weninger *et al.*, 2015), *memory-enhanced RNNs* (Weninger *et al.*, 2014a), and *deep recurrent autoencoders* (Weninger *et al.*, 2014c). Latest approaches to *speech source separation* also employ different DNN types, such as *feed-forward neural networks* (*FFNNs*; Naithani *et al.*, 2016), RNNs (Huang *et al.*, 2015; Sun *et al.*, 2017) or *end-to-end (E2E) learning* using a CNN- or RNN-autoencoder instead of the usual spectral features (Venkataramani *et al.*, 2017). Recently, *generative adversarial nets (GANs)* were found to be promising in modelling speech (Subakan and Smaragdis, 2018) and singing sources (Fan *et al.*, 2018). For the task of music source separation, it was found that both FFNNs and RNNs are suitable, achieving superior scores in the *Signal Separation Evaluation Campaign (SiSEC)* music task (Uhlich *et al.*, 2017). Latest efforts in music source separation employed *U-nets*, a CNN variant from the image processing domain (Jansson *et al.*, 2017). Moreover, a *weakly labelled data* approach has also been proposed for the task of singing voice separation (Kong *et al.*, 2017). This approach utilised information about the presence or absence of singing as given by the output of a diarisation system. Notably, despite the huge amount of publications in the field of source separation, cross-domain audio signal separation (i. e., seperation of audio sources with distinct variance in character) is still largely unexplored.

### Audio understanding

We consider audio understanding to be the task of acquiring a higher level semantic understanding of acoustic scenes, sound events, speech, and music. We consider audio understanding the task of acquiring a higher level semantic understanding of acoustic scenes, sound events, speech, and music. For this task, the aim of understanding the audio goes beyond the simple identification of speech, music, objects or events and their respective attributes. The goal, instead, should be to understand the relations between the elements of an acoustic scene. This understanding includes their relation to each other as well as their contextual meaning to a listener. For example, two individuals speaking loudly, followed by door slam and then a person crying, could be understood as a heated discussion causing emotional implications.

Unlike the field of computer vision, where considerable research has been carried out on higher-levels of semantic understanding of visual tasks (e. g., visual question answering (Agrawal *et al.*, 2017), image captioning (Xu *et al.*, 2015)), only a few works have been realised in the audio domain. One example is the recent work described in (Drossos *et al.*, 2017), in which an *encoder-decoder neural network* is used to process a sequence of Mel-band energies and to compute a sequence of words that describe a given audio segment. The already proved success of encoder-decoder sequence to seqence (S2S) architectures for structured prediction tasks such as more general audio combined with the small number of existing works applying such models to audio understanding tasks (to the best of our knowledge) creates a window of opportunity for conducting successful research in applying encoder-decoder for the above-mentioned tasks.

### Audio ontologisation

A core component of an across-audio analysis, for both interpretation and understanding of acoustic scenes, is multi-domain audio ontologisation. An ontology is a formally documented knowledge base, which provides a precise description of the concepts encompassed within a domain, with additional attributes of each concept describing possible features. Within the machine learning community, ontologisation has been widely studied and applied in the text analysis domain (Buitelaar *et al.*, 2005), human activity recognition (Hoelzl *et al.*, 2014), and for "hierarchical" image-understanding domains (Durand *et al.*, 2007). In the audio domain, however, due to the complexities of the everyday life soundscapes, most efforts have been focused on specific domains (Han *et al.*, 2010; Nakatani and Okuno, 1998).

To date, there have been scarce attempts to create complete cross-audio domain ontologisations of everyday life soundscapes. The AudioSet (Gemmeke *et al.*, 2017) by Google has been perhaps the most interesting audio ontologisation attempt to date. It offers an ontologisation of audio events and their relationships within a sub-field, i. e., classes include; music, animals, human sounds, and the corresponding dependent children are; rock, dog, and whistling. AudioSet, however, does not include descriptors of the audio (e. g., the object action, or emotion). This aspect aside, it does provide a

platform for further and deeper ontologisation by the computer audition community. Until the release of AudioSet, the majority of works in ontologisation of acoustic scenes had come from studies focusing on the ontologisation of explicit audio domains, e. g., for music genre classification (Raimond *et al.*, 2007), music emotion perception (Han *et al.*, 2010), and audio features (Allik *et al.*, 2016). Excluding AudioSet, attempts at multi-domain audio ontologisation have mainly focused on the segregation of speech and music (Nakatani and Okuno, 1998), or sound objects retrieval (Hatala *et al.*, 2004).

In order to build a basis for ontologising a domain, previous research has commonly functioned in a manual nature, developing a methodology for collaborative ontology development via data mining based visual user interfaces, such as Orange WorkFlows (known as OWLs; Hilario *et al.*, 2009). These methods create a simple "seed" of basic concepts for the ontology structure (Noy *et al.*, 2006), with further adaptations requiring huge amounts of collaborative labour, using mechanisms for carrying out discussion (e. g., polling, and moderators; Farquhar *et al.*, 1997), something which in the long run can be time-consuming and costly. In an attempt to automate the construction of an ontology ((known as ontology learning; Gotmare, 2017), there have been efforts in the field of natural language processing, for intelligent web crawling (Maedche and Staab, 2001; Ehrig and Maedche, 2003). The web offers a mass of diverse but fragmented data sources, and targets for this can include Wikipedia, YouTube, and WordNet (Gemmeke *et al.*, 2017). Such approaches use relevance computation (Zheng *et al.*, 2008), to prioritise URLs of high relevance to the data which needs to be labelled, and extract metadata from social media, e. g., comments, tags, or titles. This textual data is then clustered into groups which may provide meaning to the associated data. To create these potential clustered groupings, unsupervised learning methods for data classification have been applied in the past (Vicient *et al.*, 2013), as well as semi-supervised and active learning methods, in which categories are assigned based on the most informative instances (Gotmare, 2017).

Until this point, the deep ontologisation of a particular domain has been time-consuming, requiring a mass of human labour (even the state-of-the-art AudioSet ontology required a huge amount of manual human effort; Gemmeke *et al.*, 2017).An across-audio–domain approach will not only improve on the state-of-the-art through the inherent need for additional and more expansive audio event terminology (e. g., body acoustics, animal calls, or automotive functions), but also through more fine-grained event attributes at both the state (e. g., mood) and the trait (e. g., age) level. A starting point can be given by exploiting deep learning-based approaches for web crawling (Amiriparian *et al.*, 2017), and clustering sourced data, as well as intelligent crowdsourcing approaches to reduce the need for manual labour, in which active learning is applied to prioritise the most informative instances (Hantke *et al.*, 2017).

## TOWARDS THE NEXT GENERATION OF AUDIO INTELLIGENCE

From the above, we conclude that audio is largely being treated as a single-domain phenomenon, but the ingredients needed for a full-fledged "holistic" and likewise, an

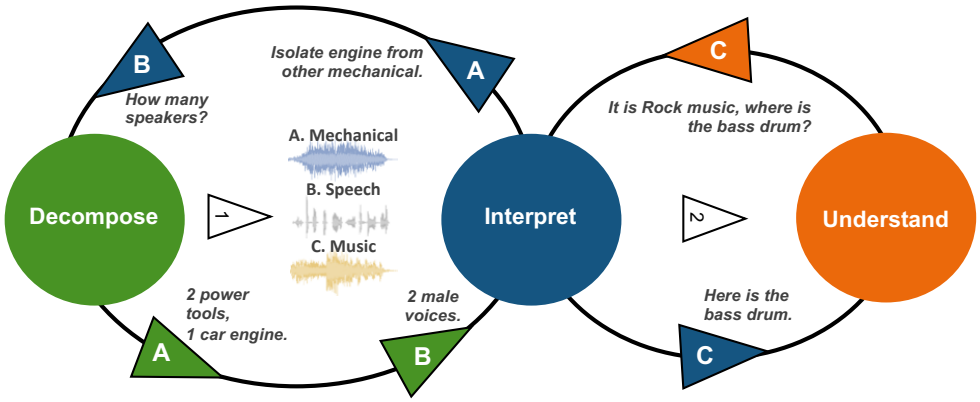


**Fig. 1:** Example for an iterative approach to decompose audio interpreting on different semantic levels of "understanding" to lead to an optimal "holistic" audio understanding. Imagine a garage with two people working on a car and listening to music as the (acoustic) scene.

arguably more "human-like" audio understanding are primarily available. In other words, one mainly needs to put the pieces of the puzzle together, and then feed a learning system with sufficient audio data. To overcome data sparseness, many approaches described in the literature use auditory and visual information in tandem to improve the understanding of video content. In Aytar *et al.* (2016), a neural network is trained on a corpus of unlabelled videos to match the representation extracted from the audio part with that extracted from the visual information by pretrained networks for object and scene classification. Facilitating such research avenues, there exist a number of video corpora, that can be used for a multimodal video understanding such as Rohrbach *et al.* (2015) and Torabi *et al.* (2015).

Figure 1 visualises a potential concept towards such holistic audio intelligence. It uses an example of an acoustic scene, as described in the introduction. The number and type of sources present in an audio signal are not known beforehand. Hence, decomposition could be modelled as an iterative process in interaction with an interpretation component, which is providing information about the signal and indicating a request for further separation, as illustrated in Figure 1. In the proposed across-audio-domain iterative decomposition solution, the first step would be to decompose speech, music, and sound and send separate signals to the interpretation component. The interpreter would be able to identify the types and then call the source separation again to decompose the signal events further. The source separation is aided by weak labels from the diarisation in this context, to know the temporal occurrences of the fractionally overlapping events. Finally, after the types of the audio have been classified by the interpretation component, these are analysed deeper with respect to states, finding that potentially parts are missing from a semantically higher perspective. This deeper analysis allows for an iterative process. Figure 2 additionally

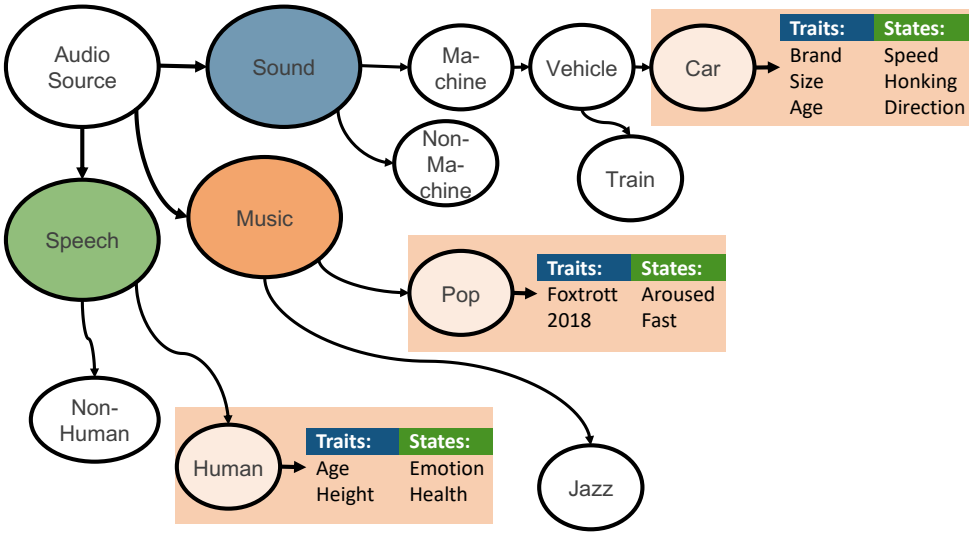


**Fig. 2:** Example of an ontology that consequently attributes audio sources states and traits – not only for speech as is the current usual state-of-literature. In this depiction we see that the audio source is decomposed into 3 sub-sources; speech, music and sound, which are then each further decomposed. For example, one of the "sound" sources is noted as being mechanical, vehicle, car, and the car is further labelled for its brand, as well as action e. g., Speed.

exemplifies audio ontologies that could suit the need for a complete and "holistic" audio understanding. Note that the concept of state and trait assignment as known from speech analysis is consequently extended to general audio sources such as sound or music – after all, sound always has a source which has certain traits and is in certain states.

## CONCLUSION

We discussed the state-of-the-art in audio intelligence focusing on audio understanding when it comes to general audio which often consists of a blend of speech and/or music and/or sound. We surveyed in nutshell components which we believe are crucial to lead to a general audio understanding including audio diarisation, source separation, understanding, and ontologisation. From this, we showed a potential approach on how to combine the pieces to lead to a more advanced form of "cross-domain" audio analysis with a rich ontology unified across the audio domains. To realise such a concept, recent deep learning methods seem well suited, such as learning weakly supervised in an end-to-end manner. Once realised, such an audio intelligence will find an abundance of potential applications from retrieval to robotics, and beyond.

## REFERENCES

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (**2017**), "VQA: Visual Question Answering," Int. J. Comput. Vis., **123**(1), 4–31.

Allik, A., Fazekas, G., and Sandler, M. B. (**2016**), "An Ontology for Audio Features," Proc. International Society for Music Information Retrieval Conference (ISMIR) (ISMIR, New York, NY), 73–79.

Amiriparian, S., Pugachevskiy, S., Cummins, N., Hantke, S., Pohjalainen, J., Keren, G., and Schuller, B. (**2017**), "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," Proc. Biannual Conference on Affective Computing and Intelligent Interaction(ACII) (San Antonio, TX), 340–345.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (**2012**), "Speaker diarization: A review of recent research," IEEE Trans. Audio Speech Lang. Process., **20**(2), 356–370.

Aytar, Y., Vondrick, C., and Torralba, A. (**2016**), "SoundNet: Learning sound representations from unlabeled video," Proc. Advances in Neural Information Processing Systems (NIPS) (MIT Press, Barcelona, Spain), 892–900.

Bredin, H. (**2017**), "TristouNet: Triplet Loss for Speaker Turn Embedding," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, New Orleans, LA), 5430–5434.

Buitelaar, P., Cimiano, P., and Magnini, B. (**2005**), *Ontology learning from text: methods, evaluation and applications* (Impacting the World of Science Press, Amsterdam, The Netherlands).

Coutinho, E., Weninger, F., Schuller, B., and Scherer, K. (**2014**), "The Munich LSTM-RNN approach to the MediaEval 2014 "Emotion in Music" Task," Proc. MediaEval Multimedia Benchmark Workshop (CEUR, Barcelona, Spain), no pagination.

Davis, K., Biddulph, R., and Balashek, S. (**1952**), "Automatic recognition of spoken digits," J. Acoust. Soc. Am., **24**(6), 637–642.

Drossos, K., Adavanne, S., and Virtanen, T. (**2017**), "Automated audio captioning with recurrent neural networks," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (IEEE, New Paltz, NY), 374–378.

Durand, N., Derivaux, S., Forestier, G., Wemmert, C., Gançarski, P., Boussaid, O., and Puissant, A. (**2007**), "Ontology-based object recognition for remote sensing image interpretation," Proc. IEEE International Conference on Tools with Artificial Intelligence (ICTAI) (IEEE, Patras, Greece), 472–479.

Ehrig, M. and Maedche, A. (**2003**), "Ontology-focused Crawling of Web Documents," Proc. ACM Symposium on Applied Computing (SAC) (ACM, Melbourne, Florida), 1174–1178.

Fan, Z., Lai, Y., and Jang, J. R. (**2018**), "SVSGAN: Singing Voice Separation Via Generative Adversarial Network," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Calgary, Canada), 726–730.

Farquhar, A., Fikes, R., and Rise, J. (**1997**), "The Ontolingua Server: A tool for

collaborative ontology construction," Int. J. Hum.-Comput. St., **46**(6), 707–727.

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., and McCree, A. (**2017**), "Speaker diarization using deep neural network embeddings," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (New Orleans, LA), 4930–4934.

Gemmeke, J., Ellis, D., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (**2017**), "Audio set: An ontology and human-labeled dataset for audio events," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, New Orleans, LA), 776–780.

Gotmare, P. (**2017**), "Methodology for Semi-Automatic Ontology Construction using Ontology learning: A Survey," IJCA Proceedings on Emerging Trends in Computin, volume ETC-2016, 1–3.

Han, B., Rho, S., Jun, S., and Hwang, E. (**2010**), "Music emotion classification and context-based music recommendation," Multimed. Tools Appl., **47**(3), 433–460.

Hansen, J. H. L. and Hasan, T. (**2015**), "Speaker Recognition by Machines and Humans: A tutorial review," IEEE Signal Process. Mag., **32**(6), 74–99.

Hantke, S., Zhang, Z., and Schuller, B. (**2017**), "Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world," Proc. INTERSPEECH (ISCA, Stockholm, Sweden), 3951–3955.

Hatala, M., Kalantari, L., Wakkary, R., and Newby, K. (**2004**), "Ontology and rule based retrieval of sound objects in augmented audio reality system for museum visitors," Proc. ACM Symposium on Applied Computing (SAC) (ACM, Nicosia, Cyprus), 1045–1050.

Hilario, M., Kalousis, A., Nguyen, P., and Woznica, A. (**2009**), "A data mining ontology for algorithm selection and meta-mining," Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) (Bled, Slovenia), 76–87.

Hinton, G., Deng, L., Yu, D., Dahl, G., rahman Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., and Sainath, T. (**2012**), "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Process. Mag., **29**(6), 82–97.

Hoelzl, G., Ferscha, A., Halbmayer, P., and Pereira, W. (**2014**), "Goal oriented smart watches for cyber physical superorganisms," Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (ACM, Seattle, WA), 1071–1076.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2015**), "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," IEEE/ACM Trans. Audio Speech Lang. Process., **23**(12), 2136–2147.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. (**2017**), "Singing voice separation with deep U-Net convolutional networks," Proc. International Society for Music Information Retrieval Conference (ISMIR) (ISMIR, Suzhou, China), 323–332.

Kong, Q., Xu, Y., Wang, W., and Plumbley, M. D. (**2017**), "Music Source Separation

using Weakly Labelled Data," Proc. International Society for Music Information Retrieval Conference (ISMIR) (Suzhou, China), no pagination.

Le Lan, G., Charlet, D., Larcher, A., and Meignier, S. (**2017**), "A Triplet Ranking-based Neural Network for Speaker Diarization and Linking," Proc. INTERSPEECH (ISCA, Stockholm, Sweden), 3572–3576.

Le Roux, J., Hershey, J. R., and Weninger, F. (**2015**), "Deep NMF for speech separation," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Brisbane, Australia), 66–70.

Maedche, A. and Staab, S. (**2001**), "Ontology learning for the semantic web," IEEE Intell. Syst., **16**(2), 72–79.

Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., and Plumbley, M. D. (**2018**), "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," IEEE/ACM Trans. Audio. Speech Lang. Process., **26**(2), 379–393.

Naithani, G., Parascandolo, G., Barker, T., Pontoppidan, N. H., and Virtanen, T. (**2016**), "Low-latency sound source separation using deep neural networks," Proc. Global Conference on Signal and Information Processing (GlobalSIP) (Washington, DC), 272–276.

Nakatani, T. and Okuno, H. G. (**1998**), "Sound ontology for computational auditory scene analysis," Proc. Conference of the Association for the Advancement of Artificial Intelligence (AAAI) (Madison, WI), 1004–1010.

Nikunen, J., Diment, A., and Virtanen, T. (**2018**), "Separation of Moving Sound Sources Using Multichannel NMF and Acoustic Tracking," IEEE/ACM Trans. Audio Speech Lang. Process., **26**(2), 281–295.

Noy, N. F., Chugh, A., Liu, W., and Musen, M. A. (**2006**), "A Framework for Ontology Evolution in Collaborative Environments," Proc. International Semantic Web Conference (ISWC) (Athens, GA), 544–555.

Ozerov, A., Févotte, C., Blouet, R., and Durrieu, J.-L. (**2011**), "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Prague, Czech Republic), 257–260.

Raimond, Y., Abdallah, S. A., Sandler, M. B., and Giasson, F. (**2007**), "The Music Ontology," Proc. International Society for Music Information Retrieval Conference (ISMIR) (Vienna, Austria), 417–422.

Reynolds, D. A. and Torres-Carrasquillo, P. (**2005**), "Approaches and applications of audio diarization," Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (IEEE, Philadelphia, PA), 953–956.

Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. (**2015**), "A dataset for Movie Description," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Boston, MA), 3202–3212.

Subakan, Y. C. and Smaragdis, P. (**2018**), "Generative Adversarial Source Separation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, Calgary, Canada), 26–30.

Sun, Y., Zhu, L., Chambers, J. A., and Naqvi, S. M. (**2017**), "Monaural source separation based on adaptive discriminative criterion in neural networks," Proc. International Conference on Digital Signal Processing (DSP) (London, UK), 1–5.

Torabi, A., Pal, C., Larochelle, H., and Courville, A. (**2015**), "Using descriptive video services to create a large data source for video annotation research," arXiv preprint arXiv:1503.01070.

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y. (**2017**), "Improving music source separation based on deep neural networks through data augmentation and network blending," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (New Orlearns, LA), 261–265.

Venkataramani, S., Casebeer, J., and Smaragdis, P. (**2017**), "Adaptive Front-ends for End-to-end Source Separation," Proc. Conference on Neural Information Processing Systems (NIPS) (Long Beach, CA), no pagination.

Vicient, C., Sánchez, D., and Moreno, A. (**2013**), "An automatic approach for ontology-based feature extraction from heterogeneous textualresources," Eng. Appl. Artif. Intel., **26**(3), 1092–1106.

Virtanen, T. (**2007**), "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," IEEE Trans. Audio Speech Lang. Process., **15**(3), 1066–1074.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., and Moreno, I. L. (**2017**), "Speaker diarization with LSTM," arXiv preprint arXiv:1609.04301.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B. (**2015**), "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," Proc. International Conference on Latent Variable Analysis and Signal Separation (Liberec, Czech Republic), 91–99.

Weninger, F., Eyben, F., and Schuller, B. (**2014**a), "Single-channel speech separation with memory-enhanced recurrent neural networks," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy), 3709–3713.

Weninger, F., Hershey, J. R., Le Roux, J., and Schuller, B. (**2014**b), "Discriminatively trained recurrent neural networks for single-channel speech separation," Proc. Global Conference on Signal and Information Processing (GlobalSIP) (Atlanta, GA), 577–581.

Weninger, F., Watanabe, S., Tachioka, Y., and Schuller, B. (**2014**c), "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (Florence, Italy), 4623–4627.

Weninger, F., Wöllmer, M., and Schuller, B. (**2012**), "Combining Bottleneck-BLSTM and Semi-Supervised Sparse NMF for Recognition of Conversational Speech in Highly Instationary Noise," Proc. INTERSPEECH (Portland, OR), 302–305.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (**2015**), "Show, Attend and Tell: Neural Image Caption Generation with

Visual Attention," Proc. International Conference on Machine Learning (ICML) (Lille, France), 2048–2057.

Zheng, H.-T., Kang, B.-Y., and Kim, H.-G. (**2008**), "An ontology-based approach to learnable focused crawling," Inf. Sci., **178**(23), 4512–4522.