Advanced Data Expoitation in Speech Analysis

An overview

Zixing Zhang, Nicholas Cummins,
and Björn Schuller

With recent advances in machine-learning techniques for automatic speech analysis (ASA)—the computerized extraction of information from speech signals—there is a greater need for high-quality, diverse, and very large amounts of data. Such data could be game-changing in terms of ASA system accuracy and robustness, enabling the extraction of feature representations or the learning of model parameters immune to confounding factors, such as acoustic variations, unrelated to the task at hand. However, many current ASA data sets do not meet the desired properties. Instead, they are often recorded under less than ideal conditions, with the corresponding labels sparse or unreliable.

In addressing these issues, this article provides a comprehensive overview of state-of-the-art ASA data exploitation techniques that have been developed to take advantage of knowledge gained from related but unlabeled or different data sources to improve the performance of a particular ASA task of interest. We first identify three primary data challenges: sparse, unreliable, and unmatched data. We then review the corresponding approaches. The conditions, advantages, and drawbacks of using a range of differing data-mining techniques are also discussed. Finally, other data challenges and potential future research directions in this field are presented.

Introduction to automatic speech analysis

ASA has long been regarded as one of the most vital areas in achieving natural and friendly human–machine interactions [1], [2]. The goal of ASA is to empower machines to automatically discern information of interest from human speech, e.g., identifying what is being said (the linguistic content), who is saying it (the speaker's identity), and how they are saying it (the paralinguistic content). More formally, typical ASA tasks in the literature include, but are not limited to,

- automatic speech recognition (ASR), which aims to extract linguistic content (e.g., words) by recognizing and translating spoken speech

- speaker identification/verification, which targets obtaining the speaker's identity from speech signals
- computational paralinguistics, which attempts to distill nonlinguistic information mainly concerning the speaker's short-term states (e.g., emotions), medium-term states (e.g., health condition and attitude), and long-term traits (e.g., personality, age, and gender) from spoken speech.

A serious obstacle to the broad application of ASA is the lack of sufficiently labeled data in terms of both quantity and quality. For example, many available computational paralinguistics corpora contain only a few hours of audio data at most [3]. Similarly for ASR, many of the world's languages are in a low-resource setting, where the electronic speech resources and linguistic expertise are lacking. According to a 2010 United Nations Educational, Scientific, and Cultural Organization report [4], approximately 2,500 languages are in danger of becoming extinct. In this scenario, it is exceptionally difficult to obtain a large-scale amount of transcribed speech data to perform reliable ASR.

> **A serious obstacle to the broad application of ASA is the lack of sufficiently labeled data in terms of both quantity and quality.**

The requirement for large-scale labeled data is not new in machine leaning. Prevailing paradigms are often conducted in a supervised manner, and a substantial increase in the amount of available training data usually brings encouraging performance improvements [5]. Because of the advancement of deep-learning technologies [6], [7], this need for data has become more compelling than ever. Deep-learning models are often designed with millions of parameters, and, if trained with insufficient amounts of data, are vulnerable to being trapped in a locally optimized minimum, resulting in overfitting to the training data [6]. When sufficiently trained, however, deep models reach unprecedented levels of performance. For example, Amodei et al. [7] utilized approximately 12,000 and 9,000 h of speech data to model English and Mandarin ASR systems, respectively, by employing deep-learning models with more than 35 million trainable parameters, achieving a performance breakthrough that exceeds the capability of even human perception. Sufficient and reliably labeled data, when available, provide the opportunity to train robust ASA models whose resulting recognition is largely invariant in the face of the abundance of acoustic variations naturally present in speech data.

Opportunities

Traditionally, tasks such as data collection and annotation have been performed by small groups of experts in a laboratory setting. This conventional work paradigm is often tedious, time consuming, and costly. However, the ongoing information and communication technologies revolution and related technologies, such as the Internet of Things (IoT) and cloud computing, are providing us with opportunities to exploit larger amounts of speech data in more effective ways than ever before.

The IoT, as a global infrastructure of the information society, is expected to offer advanced services (i.e., data collection) by interconnecting a wide variety of contemporary recording devices, such as smartphones, wearable devices, and tablets. Furthermore, as these devices often have microphones, social media apps, and Internet connectivity, they can be considered distributed sensors or entryways for speech collection and processing. Thus, the advance of Internet technologies and the ubiquity of smart devices can drastically reduce the cost and time associated with collecting and processing speech data.

Cloud computing, or Internet-based computing, is expected to provide an on-demand computing resource. Thus, it gives an opportunity to store, access, and analyze the volume of speech data generated by the distributed devices mentioned previously. Cloud computing has been shown not only to minimize the costs associated with an ever-increasing demand for greater computational resources but also to reduce the cost associated with infrastructure maintenance and user access. Motivated by these advantages, most major speech technology providers have already shifted their primary research and application attention from embedded systems to cloud computing platforms.

Generalized automatic speech analysis: Problem statement and notation

The aforementioned technologies provide great potential to generate and process a large amount of speech data. However, there are three main challenges—data sparsity, unreliability, and nonmatching (Figure 1)—that limit the dissemination of these data in research and industry. Before formally defining these challenges, we first overview the generalized mathematical problem statement and notation commonly used in both ASA and throughout the remainder of this article.

First, let us define a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ that comprises a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where X denotes a set of feature vectors, i.e., $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}$; while $P(X)$ indicates the distribution of X in $\mathcal{X}$. In the case that each feature vector $\mathbf{x}$ consists of d attributes, i.e., $\mathbf{x} = \{x_1, \ldots, x_d\}$, $\mathcal{X}$ is a d-dimensional space. The most commonly used feature space $\mathcal{X}$ for ASA is arguably the Mel-frequency cepstral coefficients (MFCCs) that are extracted via filtering a speech frame by a bank of nonlinear bandpass filters (Mel filters) whose frequency response is based on the cochlea of the human auditory system [8]. Other exemplary feature spaces include the *i*-vector representation often used for speaker identification/verification [9], and mixed brute force feature representations, such as the broadly used ComParE feature set, which contains 6,373 static features (i.e., statistical functionals including mean and variance) of low-level descriptor (LLD) contours (i.e., MFCCs) often used in tasks such as recognition of emotion from speech [10].

We further define a generic ASA task $\mathcal{F} = \{\mathcal{Y}, f(\cdot)\}$ that consists of a label space $\mathcal{Y}$ and a predictive function $f(\cdot)$ (or a conditional distribution $P(Y \mid X)$). The goal of this task is to build an effective and robust predictive function $f(\cdot)$ that is capable of learning transformation rules from the feature space

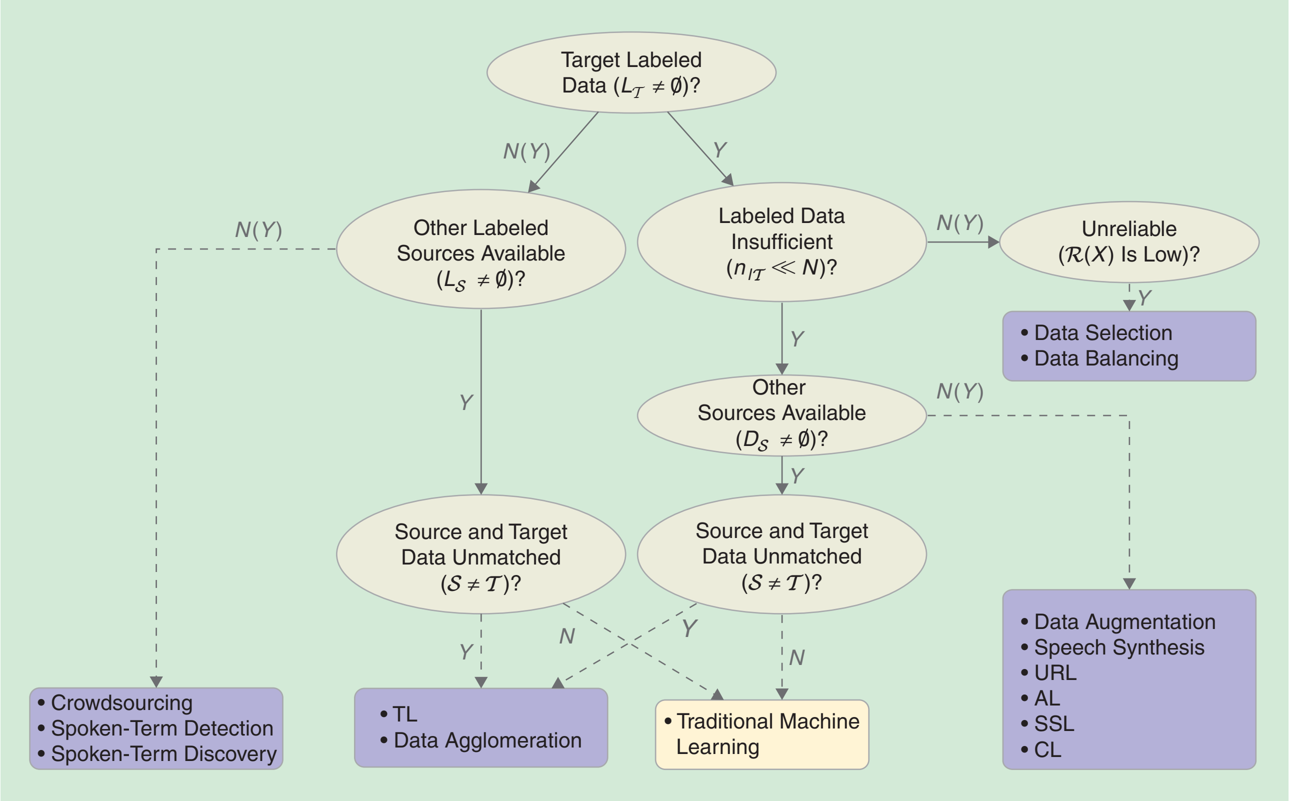
Target Labeled Data ($L_{\mathcal{T}} \neq \emptyset$)?
N(Y)
Y
Other Labeled Sources Available ($L_{\mathcal{S}} \neq \emptyset$)?
N(Y)
Labeled Data Insufficient ($n_{l\mathcal{T}} \ll N$)?
N(Y)
Unreliable ($\mathcal{R}(X)$ Is Low)?
Y
• Data Selection
• Data Balancing
Y
Y
Other Sources Available ($D_{\mathcal{S}} \neq \emptyset$)?
N(Y)
Y
Source and Target Data Unmatched ($\mathcal{S} \neq \mathcal{T}$)?
Source and Target Data Unmatched ($\mathcal{S} \neq \mathcal{T}$)?
Y
N
Y
N
• Crowdsourcing
• Spoken-Term Detection
• Spoken-Term Discovery
• TL
• Data Agglomeration
• Traditional Machine Learning
• Data Augmentation
• Speech Synthesis
• URL
• AL
• SSL
• CL

FIGURE 1. A taxonomic overview of the three main data challenges associated with ASA and their potential solutions as discussed in this article. Note that $N(Y)$ denotes no or yes, which indicates the possible combination of techniques. TL: transfer learning; AL: active learning; SSL: semisupervised learning; CL: cooperative learning; URL: unsupervised representation learning.

$\mathcal{X}$ to the label space $\mathcal{Y}$, i.e., $\mathcal{X} \xrightarrow{f(\cdot)} \mathcal{Y}$. Then, when given a test sample, it maps this feature vector $\mathbf{x}_*$ into a specific label y_*, i.e.,

$$y_* = f(\mathbf{x}_*), \tag{1}$$

where $\mathbf{x}_* \in \mathcal{X}$ and $y_* \in \mathcal{Y}$. As an example, when performing ASR, $y_* \in \mathcal{Y}$ denotes a phoneme or a word; $f(\cdot)$ is then trained to predict a phoneme or a word from, e.g., MFCCs. In speaker identification/verification, $y_* \in \mathcal{Y}$ denotes a speaker identity; the $f(\cdot)$ is trained to predict speaker identity, e.g., from *i*-vectors. Similarly, in speech emotion recognition, $y_* \in \mathcal{Y}$ denotes an emotional state, and $f(\cdot)$ is trained to recognize the emotional state, e.g., from high-dimensional statistical features.

Given a domain $\mathcal{D}$ and a task $\mathcal{F}$, we define D to denote a speech database. As the majority of available pattern recognition approaches are supervised paradigms [the input and expected output for $f(\cdot)$ are provided during training]. A database is normally given by two parts: the feature vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}$ and the corresponding labels $Y = \{y_1, \ldots, y_n\} \in \mathcal{Y}$. However, in real life, the labels y_i are often only partially provided (or not even provided) because of the difficulty of labeling. In this case, we denote the labeled data partition as $L = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_l}, y_{n_l})\}$ and the unlabeled data partition as $U = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_u}\}$, where n_l and n_u are the total number of labeled and unlabeled instances, respectively. In this sense,

$$D = L \cup U, \tag{2}$$

and $n = n_l + n_u$.

Furthermore, we define the domain for the target task to be the target domain $\mathcal{T}$. The data in this domain might be insufficient for training an effective and robust prediction function $f(\cdot)$. For example, when performing ASR on a low-resource language, $\mathcal{T}$ could be a language such as Assamese, Bengali, Haitian, Lao, Pashto, Tamil, Tagalog, Xitsonga, or Zulu [11]. In this case, we define other domains from which data could be leveraged for the target task as source domains $\mathcal{S}$. For example, for low-resource ASR, one $\mathcal{S}$ could be a high-resource language such as English or Mandarin [11]). According to (2), then

$$D_{\mathcal{T}} = L_{\mathcal{T}} \cup U_{\mathcal{T}}, \tag{3}$$

and

$$D_{\mathcal{S}} = L_{\mathcal{S}} \cup U_{\mathcal{S}}. \tag{4}$$

In this article, we use the term *data* interchangeably with *instance*, *turn*, *record*, *utterance*, *segment*, *sample*, or *example*. Similarly, the term *annotator* is interchangeable with *evaluator*, *transcriber*, *labeler*, or *translator*; and the word *annotation* is used to denote any labeling task, i.e., transcription for ASR or labeling the emotion or other speaker states and traits associated with an utterance.

Data challenges

This section offers a detailed overview into the data sparsity, data unreliability, and unmatched data challenges. Techniques to adequately cope with these challenges will play an essential role in the development of the next generation of reliable ASA systems.

Sparse data challenge

While there is an abundance of raw speech data, the corresponding annotations needed for many ASA tasks are often scarce (i.e., $L_{\mathcal{T}} \neq \emptyset$, but $n_l \ll N$, where $\emptyset$ denotes the empty set and N is a required number), or nonexistent (i.e., $L_{\mathcal{T}} = \emptyset$). For example, outside of speech recognition tasks on a handful of widely used languages (e.g., English and Mandarin), the labels needed to conduct ASR on other languages are particularly scarce (see the Intelligence Advanced Research Projects Activity [IARPA] Babel project [11]). Similarly, most databases available for computational paralinguistics tasks, such as emotion recognition and personality analysis, may contain 5 h of labeled data at most [12], [13], which is insufficient for building highly robust models.

> **Thanks to the pervasive sensing opportunities offered by smart devices and social media, the gathering of speech data has become a somewhat easier task.**

However, thanks to the pervasive sensing opportunities offered by smart devices and social media, the gathering of speech data has become a somewhat easier task. For example, it is reported that some 500 h of video content is being uploaded to the video-sharing website YouTube every minute [14]. Nonetheless, labeling these data demands huge amounts of expert manual labor, which is regarded as being prohibitively expensive and time consuming. Taking speech transcription as an example, it can take up to approximately 6 h to accurately transcribe 1 h of speech at an average price of US$150/h [15], [16]. While a few Internet giants (e.g., Amazon, Google, and Microsoft) have the capability of obtaining many thousands of annotated speech data for ASA tasks, such as speech recognition, these labeled data are, however, rarely made freely available to interested research groups.

If $D_{\mathcal{T}}$ does not contain any labeled data, i.e., $L_{\mathcal{T}} = \emptyset$, a naïve solution is manual annotation. An efficient way to do this is using a crowdsourcing platform, which is an Internet-based system that utilizes a large group of individuals to perform a common service. Alternatively, spoken-term detection/discovery can be considered as a means of detecting predefined patterns in the data or discovering unknown patterns there.

If $D_{\mathcal{T}}$ contains some labeled data, i.e., $L_{\mathcal{T}} \neq \emptyset$, it is then necessary to assess whether or not the available labeled data are sufficient in terms of quantity and diversity to develop a robust model. If the data are found to be insufficient, data augmentation approaches, which seek to enrich the number and variety of existing labeled speech data, might be an appropriate option. A further option is the use of speech synthesis to automatically generate data with predefined labels. If a large scale of unlabeled data are available, i.e., $U_{\mathcal{T}} \neq \emptyset$, alternative solutions could include unsupervised representation learning (URL), semisupervised learning (SSL), active learning (AL), and cooperative learning (CL). These techniques are becoming prominent paradigms to efficiently leverage massive unlabeled data via a small amount of labeled seed data [17].

Unreliable data challenge

This is the scenario in which the total amount of speech data is large, but the data reliability is low. Data sets collected in real-life settings, and even many collected in controlled laboratory settings, are susceptible to a range of problems, such as distortion by environmental noises, recording devices, or interfering speakers [18]. Besides, the associated annotations may be unreliable because of mistakes or high uncertainty among multiple annotators [18]. Furthermore, in many cases, the distribution of collected speech data can be highly unbalanced over the classes of interest. All these factors can give rise to noisy and unreliable data, leading to nontrivial difficulties when training models [18], [19].

Additionally, the reliability of the labeled data should be evaluated in terms of properties such as acoustic quality, annotation certainty, and data balance degree. Poor data quality has frequently shown its detrimental effect on system performance. In this scenario, data selection should be considered for eliminating the noisy, unrelated, and unreliably labeled data or data balancing for balancing the data distribution.

Unmatched data challenge

This is the situation where data from a target domain $\mathcal{T}$ are not sufficient or reliable enough to train a robust model for a task of interest. However, as previously discussed, there are often data from a source domain $\mathcal{S}$ that are easy to obtain and somehow related to the target data. This motivates researchers to explore leveraging source domain data to aid the target ASA task. For example, one of the goals of the IARPA Babel project is to utilize the available and large-scale speech data in, e.g., the English language for speech recognition in low-resource languages. Nevertheless, in many real-world applications, the source and target domains are often highly unmatched in respect to acoustic signal conditions, speakers, tasks, or even recording devices [20]. These mismatches lead to a marked performance degradation of the analysis in such models in real-life settings [20], [21].

Mathematically, the source domain can differ from the target domain (i.e., $\mathcal{S} \neq \mathcal{T}$) in terms of 1) modalities, i.e., $\mathcal{X}_{\mathcal{S}} \neq \mathcal{X}_{\mathcal{T}}$ (this case is considered out of the scope of this article, which is focused only on speech), 2) marginal probability distributions, i.e., $P(X_{\mathcal{S}}) \neq P(X_{\mathcal{T}})$, 3) label spaces, i.e., $\mathcal{Y}_{\mathcal{S}} \neq \mathcal{Y}_{\mathcal{T}}$, and/or 4) conditional probability distributions, i.e.,

$P(Y_{\mathcal{S}}|X_{\mathcal{S}}) \neq P(Y_{\mathcal{T}}|X_{\mathcal{T}})$. A more in-depth explanation of these discrepancies can be found in [21].

An idealized solution to mitigate these differences is to obtain access to all possible variations by acquiring data on a massive scale. However, it is either practically impossible to anticipate all variations or such data would require exhaustive annotation. In such unmatched scenarios, transfer learning (TL) [21] is regarded to be a highly promising technique to take advantage of the knowledge from the source domain for the target domain.

Finally, it is important to note that all of the aforementioned techniques for each challenge can be performed either individually or jointly. This is illustrated in Figure 1, where possible combinations that can occur are indicated through the use of the N(Y) symbol, which denotes *no* or *yes*. For example, crowdsourcing can be used no matter whether the labeled target data are available or not. Likewise, AL can be executed on either unlabeled target data or unlabeled source data. All of the key techniques mentioned in this section are reviewed in detail in the following sections.

Another emerging trend for crowdsourcing is the gamification of the service, which is used to introduce a sense of fun into what are often simple and recurring tasks.

Contributions of this article

The literature shows a few surveys relevant to the topic of this article. Deng et al. [22] offered a comprehensive overview of machine-learning paradigms for speech recognition systems. Wang et al. [20] provided a TL survey for speech and language processing, drawing the conclusion that TL has the potential to overcome the data-mismatch challenge. None of these surveys, however, perform a complete analysis of the sparse, unreliable, and unmatched data challenges or provide a comprehensive overview of the corresponding approaches.

Extending from a previous abstract [12], this article is the first to offer a thorough and in-depth overview of the most prominent and state-of-the-art techniques in this direction, including crowdsourcing for efficient data labeling; spoken-term detection/discovery to facilitate learning when there are no labeled data; data augmentation, speech synthesis, URL, SSL, AL, and CL to enable learning when only a limited amount of labeled resources are available; data selection and balancing techniques to facilitate learning from unreliable or unbalanced resources; and TL and data agglomeration to learn unmatched resources.

Rather than simply enumerating a list of associated papers and techniques, the focus of this article is on the analysis of the various data conditions and on how to better explore data under the different conditions. In doing this, ASA researchers and developers, new and established, can profit from the approaches introduced and discussed for the aforementioned applications.

Efficient data labeling: Crowdsourcing

The most straightforward solution to address a shortage of labeled data is to organize a group of workers (i.e., annotators) to perform the required annotations. By doing this, we create a new or additional labeled data set L_{cs}, and then ASA models can learn from the increased labeled data set $L' = L \cup L_{\text{cs}}$. Manual annotation is, however, costly in terms of time and money. Therefore, strategies to reduce these costs are of particular importance.

Crowdsourcing is one method to gather the needed data in a cost-efficient manner. In crowdsourcing, human intelligence tasks (HITs) such as data annotation are distributed via the Internet to a large number of potential workers (annotators). The users perform the tasks for usually low compensation. The assumption behind crowdsourcing is that the use of nonexperts is less onerous and more rapid than the use of experts. Furthermore, the aggregated opinion of many nonexperts has been shown to approach the quality of the opinion offered by comparatively fewer experts [15], [23], [24].

Popular crowdsourcing platforms include Amazon Mechanical Turk (MTurk), CrowdFlower, and Crowdee. MTurk is likely the most popular crowdsourcing platform for ASA-related tasks. While MTurk provides access to a larger number of potential annotators, it is considered relatively expensive when compared to other platforms [15]. The CrowdFlower platform is steadily increasing its market share. When compared to Mturk, it provides customers with a steady number of contributors and has a higher degree of quality control. An emerging trend, as implemented by Crowdee, involves moving the platform from the web to a mobile platform. Participants associated with this platform have the potential to undertake a task at any time and place.

Another emerging trend for crowdsourcing is the gamification of the service, which is used to introduce a sense of fun into what are often simple and recurring tasks. This is also interesting from an ethical point of view, aiming to improve working conditions of crowd workers. The iHEARu-play platform, for example, offers annotators a chance to perform labeling, or prompted recording tasks, in return for scores and prizes, which are computed on the correctness and workload of their annotations [25].

Generally, the procedure of crowdsourcing speech resources can be broken into four stages. The first step is to define the project parameters, such as an appropriate platform, quality control strategy, budget, and time scale. The second step is to prepare the data. The third step is to distribute tasks. This generally involves splitting the whole task into many small units and then assigning each unit to several annotators. The final step is to aggregate and evaluate the resources (e.g., speech data or annotations).

For speech processing, crowdsourcing has been widely employed for a range of tasks, including speech data collection/acquisition, speech annotation, speech perception, assessment of speech synthesis, and dialog system evaluation [15], [26]. With particular respect to speech annotation, many studies have shown crowdsourcing's benefit in terms of both increased transcription quality and decreased costs. For example, in [27], the authors proposed a two-stage approach to transcribe speech via a crowdsourcing platform (i.e., microworkers). Specifically, the utterances that were labeled with the lowest agreement level among annotators would be selected for a second-stage translation. In doing this, more than 250,000 utterances (156 h) of spoken dialog from real callers were translated, being of comparable quantity to the same corpora labeled by experts but at considerably less cost. Similar work has been presented for the transcription of meeting data [24], addressing the business name queries from a publicly accessible telephone directory service [16], and labeling the emotional state of speakers [28]. All of these works show that crowdsourcing is a relatively affordable and efficient way to address the task of speech annotation, compared with conventional methods.

Crowdsourcing is a relatively affordable and efficient way to address the task of speech annotation, compared with conventional methods.

Despite the advantages, controlling the quality of the labels is important to ensure they are as reliable as those given by experts. In this regard, quality control measures are required. A range of quality control mechanisms have been proposed in the literature, which can be grouped into one of the following five categories:

1) *Worker filter*: This mechanism evaluates annotation quality through the use of control questions (a question with a restricted answer set) and filters out inappropriate annotations.
2) *Intraworker*: The reliability of an annotator can be evaluated by the consistency of the response to the same question asked multiple times. Alternatively, this could be established by a self-confidence value chosen by the annotator [27].
3) *Interworker*: Normally, a gold standard is calculated via techniques such as majority voting, using responses from a multitude of annotators. The quality of an individual annotator can then be evaluated by calculating the response dissimilarity to the gold standard. This method is, of course, susceptible to the risk that the majority results are wrong.
4) *Gold-standard comparisons*: This is a particular case of the interworker mechanism, where the gold standard is provided by trustworthy experts. This mechanism has been shown to be effective in eliminating intentionally malicious annotators, albeit at the cost of expert intervention [27], [29].
5) *Third-party review*: Here, quality control is carried out by a third party, e.g., another independent crowdsourcing task [30], or by the output of an intelligent system [16], [27]. However, this requires extra quality evaluation or computational costs.

Learning from no labeled resources

This section discusses paradigms suitable for the extreme operating scenario where no labeled data are available, i.e., $L = \emptyset$, and $D = U$. In this scenario, techniques such as spoken-term detection and spoken-term discovery, or related methods of targeted detection of speech-related information and phenomena of interest and according discovery in the sense of novelty detection can be used to identify salient information (i.e., patterns) directly from an unlabeled data set without any manual intervention. The premise of these techniques can be thought of as analogous to infant language acquisition, i.e., the learning of linguistic information from the raw speech of an unknown language during the first few years of an infant's life. The two techniques (i.e., targeted detection and novelty discovery) are distinguished by whether, e.g., spoken terms have been previously identified (spoken-term detection) or not (spoken-term discovery). Next, we focus on terms; however, similar methods can be applied to retrieve speech related to other phenomena of interest.

Spoken-term detection

The goal of spoken-term detection is to retrieve a set of occurrences from a speech repository for given acoustic queries or terms (normally spoken words or phrases). Compared with conventional speech recognition approaches, spoken-term detection offers the capability to detect corresponding patterns from speech in the absence of any text information.

The predominant spoken-term detection methods involve template-based acoustic models and typically rely on dynamic time warping (DTW) [31]. Specifically, they search for the predefined terms in a lattice. In a no-labeled-resource scenario, DTW has been shown to be an effective way to find the matched patterns [31]. Nevertheless, DTW alignment requires substantial computational resources to compare segments [32], [33]. Tackling this runtime-scalability problem is an active and ongoing research direction [32], [34]. Key approaches proposed in the literature include information retrieval-based DTW [35]. This approach first estimates the regions of an utterance that are more likely to contain the spoken query and then uses a standard DTW to find the exact start and end times of each pattern. This approach was further extended in [34] via the introduction of a hierarchical k-means clustering, contributing to a substantial speedup when compared with classic DTW.

An alternative approach is to embed the arbitrary-length segments into fixed-dimensional spaces [32]. This technique greatly reduces the computational load without any performance compromise. Following this idea, the novel framework of audio Word2Vec was recently proposed [36]. Audio Word2Vec uses a sequence-to-sequence autoencoder [a neural network (NN) commonly used as an unsupervised learning algorithm; for more details, see the "Deep Belief Networks and Stacked Autoencoders" section] to represent any arbitrary-length audio segment as a fixed-length vector. This framework was determined to outperform conventional DTW-based approaches at substantially lower computational requirements [36].

Spoken-term discovery

Spoken-term discovery, also known as *spoken-term indexing*, is the task of searching potentially large, untranscribed speech collections for recurring words and phrases without using any language-specific resources other than the collection itself [37]. Specifically, spoken-term discovery differs from spoken-term detection in that spoken-term discovery systems automatically find an inventory of lexical units (words or phrases) without being given any user-specific terms. Furthermore, spoken-term discovery is distinct from conventional ASR systems, where a lexicon is always specified.

Data augmentation artificially generates more data by transforming existing speech samples using certain transformations that preserve the original class labels and speech content.

Typically, spoken-term discovery consists of three steps [13]: 1) pairwise matching, 2) clustering, and 3) parsing. The aim of pairwise matching is to identify pairs of segments, taken from unique continuous spoken utterances, that have high acoustic similarity. Similar to spoken-term detection, the dominant techniques in this step are based on DTW.

The discovered segments are then clustered into classes (indices) that correspond to a set of likely words and phrases present in the data. Typically, an abstract adjacency graph [31] is used to represent the relationship between all of the segmented pairs. The nodes of this graph correspond to the locations in time of the segments, and its edges correspond to the measures of similarity between those time indexes. A predefined threshold is then applied to the edge weights, which results in clusters of highly connected nodes. While the edge thresholding is regarded as the de facto clustering method for spoken-term discovery, there is a range of fast and efficient algorithms for automatic graphic clustering that could be applied. For example, the work in [31] utilized the Newman algorithm, which first removes all edges and then merges potential groups together in a greedy fashion by adding edges back to the graph.

Finally, the discovered speech segments are used to parse the utterances. The identification of the segment (term) boundaries is challenging; the alignment segments are often overlapping in a particular node, and the ending times of their respective time intervals can differ. A straightforward solution for this issue is to calculate the average start and ending times for all of the alignment segments belonging to one node [31].

While considerable advances have been made for fully unsupervised speech processing, the majority of studies are limited to small-size data sets. Studies have shown that performance is dramatically degraded when facing a large data set [26] or a large variety of speakers [38]. However, this approach is still quite attractive for many low-resource ASA tasks, e.g., early language acquisition.

Learning from limited labeled resources

Rather than starting with a completely unlabeled data set, we are often in the better situation of having a limited number of labeled resources, i.e., some few and expensive labeled speech data exist $L \neq \emptyset$, while $n_l \ll N$, where N denotes an opportune number of annotations. In this scenario, a range of other techniques besides the aforementioned no-labeled-resource methods can be utilized. These are generally implemented in one of two ways: 1) increasing the size and diversity of the existing labeled data by means of manually modifying the speech variations (i.e., data augmentation) or artificially generating new speech with predefined labels (i.e., speech synthesis) or 2) the efficient leveraging of information gained from big unlabeled data, through a priori knowledge of the labeled data. Typical techniques here include URL, SSL, AL, and CL. In the following text, each of these techniques is discussed in detail, with key contributions from the literature summarized in Table 1.

Data augmentation

Data augmentation artificially generates more data by transforming existing speech samples using certain transformations that preserve the original class labels and speech content. By taking this approach, an augmented data set L_{aug} is obtained from the original data set L, i.e., $L_{\text{aug}} = AUG(L)$, which is then added to an updated labeled data set $L' = L \cup L_{\text{aug}}$. The popularity of data augmentation is indeed highly relevant to the ongoing development of deep learning, the success of which strongly depends on having large amounts of training data. Many studies have reported that training on data of limited quantity and variety leads to a failure of deep-learning systems owing to factors such as overfitting [6].

Variations in speech data are strongly influenced by numerous factors, such as the speaker's age, gender, and cultural background, and even the content of the background noise. Data augmentation techniques, through a series of transformations (perturbations), allow us to artificially increase both the quantity and variations present in some training data, consequently improving the generalizability of the classifiers trained on this data. Conventional data augmentation approaches mainly involve artificially adding noise of various types, including convolutional noise, and levels to the original training speech for training a noise-robust acoustic model in multiple acoustic conditions [39].

Recently, research efforts have focused on using more complex perturbation approaches, such as vocal tract length perturbation (VTLP) [40], or stochastic feature mapping (SFM) [41]. In VTLP, an alternate replica of an utterance is created by distorting its spectrum [40]. First, Mel-filter banks are applied over the spectrum. Then, the center frequencies (f) of all of the filter banks are mapped to new frequencies (f') by employing a warping procedure:

$$f' = f \cdot \phi(\alpha), \tag{5}$$

Table 1. Selected data-exploitation studies on the limited labeled speech resource.

Publications	Types	Approaches	Models	Applications	Databases and Languages
Weng et al. 2014 [39]	DAU	Adding noise	Recurrent DNN	ASR	WSJ0 (En)
Amodei et al. 2015 [7]	DAU	Adding noise	CNN, DNN, CTC	ASR	WSJ0 (En), Switchboard (En), Fisher (En), Baidu (En, Ma), LibrisSpeech (En)
Jaitly and Hinton 2013 [40]	DAU	VTLP	DNN, CNN	ASR	TIMIT (En)
Cui et al. 2015 [41]	DAU	VTLP, SFM	DNN, CNN	ASR, KWS	IARPA Babel program (As, Ha)
Tüske et al. 2014 [42]	DAU	VTLP	BN-MLP	ASR, KWS	IARPA Babel program (five lang.)
Ko et al. 2015 [43]	DAU	Tempo-/speed based	Time Delay NN	ASR	Switchboard (En), Gale database (Ma), LibriSpeech (En), Tedlium (En)
Peddinti et al. 2015 [44]	DAU	Volume based	Time Delay NN	ASR	Switchboard (En)
Milde and Biemann 2015 [45]	DAU	Pitch based	CNN	Eating condition classification	iHEARu-EAT corpus (En)
Schuller et al. 2012 [46]	SS	Waveform-based	SVM	ER	Two synthesized + eight human corpora
Gales et al. 2009 [47]	SS	Parameter-based	SVM, HMM	ASR	WSJ Corpus (En)
Dahl et al. 2012 [51]	URL	DBNs	DBNs	ASR	Business Search Dataset (En)
Seide et al. 2011 [64]	URL	DBNs	DBNs	ASR	Switchboard-I (En)
Deng et al. 2010 [54]	URL	SAEs and DBNs	SAEs and DBNs	Speech coding	TIMIT (En)
Mohamed et al. 2012 [65]	URL	DBNs	DBNs	ASR	TIMIT (En)
Lei et al. 2014 [66]	URL	DNNs	DNNs	Speaker recognition	NIST SRE'12 (En)
Liu et al. 2014 [67]	URL	DBNs	DBNs	Speaker identification	NIST 2005 SRE (En)
Stuhlsatz 2011 [68]	URL	DNNs	DNNs	ER	Nine emotional corpora
Sánchez-Gutiérrez et al. 2014 [69]	URL	DBNs	DBNs	ER	Spanish emotional speech database (Sp)
Kim et al. 2013 [70]	URL	DBNs	DBNs	Audiovisual ER	IEMOCAP (En)
Hau and Chen 2011 [57]	URL	Deep CNNs	Deep CNNs	Speaker/gender identification Phone classification	TIMIT (En)
Lee et al. 2009 [58]	URL	Convolutional DBNs	Convolutional DBNs	Speaker/gender identification Phone/music classification	TIMIT (En), music data
Kemp and Waibel 1999 [71]	SSL	Self-training	GMM–HMM	ASR	View4You broadcast news database (Ge)
Wessel and Ney 2005 [72]	SSL	Self-training	HMM	ASR	BROADCAST NEWS96/7 corpora (En)
Fazakis et al. 2015 [73]	SSL	Self-training	NB, SVM, LR	Speaker identification	CHAINS Corpus (En)
Hsiao et al. 2013 [74]	SSL	Self-training	MLP	KWS	IARPA Babel Program (Tu, Vi)
Thomas et al. 2013 [75]	SSL	Self-training	DNN	ASR	Callhome Corpora (En, Ge, Sp)
Zhang et al. 2013 [76]	SSL	Cotraining	SVM	Emotion/sleeping/age/gender classification	Six emotional corpora
Cui et al. 2012 [77]	SSL	Multiview learning	RDT, HMM	ASR	Broadcast News corpus (En)
Liu and Kirchhoff 2016 [78]	SSL	Graph-based learning	DNN	ASR	Switchboard (En), DARPA RM (En)
Riccardi and Hakkani-Tür 2005 [79]	AL	Uncertainty sampling	HMM	ASR	"How May I Help You?" database (En)
Varadarajan et al. 2009 [80]	AL	Uncertainty sampling	HMM	ASR	Directory assistance data (En)
Fraga-Silva et al. 2015 [81]	AL	Uncertainty sampling	GMM–HMM	ASR, KWS	IARPA Babel Program (six languages)

(*continued*)

Table 1. Selected data-exploitation studies on the limited labeled speech resource. *(continued)*

Publications	Types	Approaches	Models	Applications	Databases and Languages
Hamanaka et al. 2010 [82]	AL	Query by committee	GMM–HMM	ASR	Corpus of Spontaneous Japanese (Ja)
Zhang and Schuller 2012 [83]	AL	Meta query	SVM	ER	FAU AEC (Ge)
Zhang et al. 2015 [84]	AL	Meta query	SVM	ER	FAU AEC (Ge)
Riccardi and Hakkani-Tür 2003 [85]	CL	Confidence score	HMM	ASR	"How May I Help You?" database (En)
Yu et al. 2010 [86]	CL	Confidence score	HMM	ASR	Broadcast Conv. and News corpora (Ma)
Zhang et al. 2015 [17]	CL	Confidence score	SVM	ER	FAU AEC (Ge), SUSAS (En)
Yu et al. 2010 [87]	CL	Global-entropy based	HMM	ASR	Directory assistance data (En)

BN: Bayes network; CTC: connectionist temporal classification; NB: naive Bayes; LR: logistic regression; RDT: randomized decision making; DAU: data augmentation; SS: speech synthesis; ER: emotion recognition; As/Da/En/Fr/Ge/Ha/Ja/Ma/Sp/Tu/Vi/Xi/Zu: Assamese/Danish/English/French/German/Haitian Creole/Japanese/Mandarin/Spanish/Turkish/Vietnamese/Xitsonga/Zulu.

where α, the wrapping factor, is randomly chosen from $[0.9, 1.1]$. The results presented in [40] indicate that, in terms of the phone error rate, deep networks trained on a VTLP-augmented version of a small database can outperform the deep networks trained on the original data set. Based on that work, a deterministic perturbation (i.e., α changes in the range of warping factors with a fixed step) rather than a random perturbation was proposed and investigated [42].

SFM, inspired by voice conversion paradigms, seeks to utilize the acoustic-feature-space relationship among speakers when augmenting a data set [41]. Specifically, it augments training utterances by statistically converting one speaker's speech data to another's using

$$\mathbf{x}' = \mathbf{x} \cdot \mathcal{M}, \tag{6}$$

where $\mathcal{M}$ is a transformation matrix of the feature spaces between two speakers. The experimental results given in [41] show that SFM offers improved performance over VTLP on both ASR and keyword spotting (KWS) tasks.

Other data augmentation approaches include tempo-based, speed-based, and volume-based perturbations [43]. Tempo-based perturbation modifies the speech tempo while retaining the pitch and the spectral envelope. Speed-based perturbation varies the speech speed by resampling, whereas volume-based perturbation changes the amplitude of signals.

While data augmentation approaches have frequently been effective in ASR tasks [7], [44], this has not proved to be as much the case in other ASA tasks, particularly in computational paralinguistics [45]. A potential reason for this might be that the detection of speaker states and traits (e.g., emotion, age, and gender) is more sensitive to changes in speech variation. Therefore, training on inappropriately transformed speech would lead to a worse model. Emotion, for example, is known to be related to the speech tempo; speech with faster tempo is inclined to be recognized as higher arousal in emotion recognition, so changing the associated speech tempo from fast to slow would potentially lead to badly labeled training data.

Continued research efforts being undertaken to distinguish features that are task specific or task invariant could help facilitate the application of data augmentation to other speech analysis tasks. In addition, most recent applications of data augmentation are performed for deep learning [7]. The effectiveness of these techniques on shallow discriminative or generative models is yet to be established.

Speech synthesis

Similar to data augmentation, the speech synthesis approach aims to synthesize additional labeled data, i.e., $L_{\text{syn}} = SS(L)$, such that the new labeled data set L' is updated by $L' = L \cup L_{\text{syn}}$. Theoretically, speech synthesis can produce an infinite amount of labeled data via altering speech content or modifying the parameters of a speech synthesizer. However, as the parameters of the synthesizers have a limited range, the simulated speech data often face the problem of limited variations. This can consequently result in the overfitting issue when training models. Combining the synthesized speech data with natural instances has been shown to help minimize this overfitting issue [46]. For emotion recognition in speech, it has been shown that systems trained on synthesized speech (the test data was natural speech) can deliver competitive performance when compared to equivalent systems trained on natural speech [46]. In this article, two synthesizers rendering emotional speech—Emofilt and Mbrola—were utilized to artificially generate speech colored with predefined emotions [46].

Rather than directly synthesizing waveforms, an alternative is generating parameterized speech that can be used directly for training a discriminative classifier. Gales et al. [47] used a hidden Markov model (HMM)-based statistical synthesis to generate missing words in a training set, when building word-based support vector machines (SVMs) for ASR. The results presented indicate that this HMM-based synthesis approach was able to yield gains over the baseline. Inspired by the success of deep learning, an emerging research trend is to use NNs rather than HMMs to generate speech samples [48], which may also mature in terms of the variation of synthesized speaker states and traits.

Unsupervised representation learning

In contrast to data augmentation and speech synthesis, URL techniques attempt to leverage massive unlabeled data, rather than sparsely labeled data. URL is closely related to the pretraining process of deep learning, which aims to learn the underlying representations $\mathbf{x}'$ embedded in speech signals via multiple unsupervised transformations, i.e., $\mathbf{x}' \leftarrow URL(\mathbf{x})$, where $\mathbf{x} \in D = L \cup U$. To train a recognition model for a specific task, the pretrained model is then updated in a supervised manner via a small amount of labeled data. This step is generally referred to as *fine-tuning* or *discriminative learning*.

A typical model structure for URL is often composed of multiple processing layers of NNs for linear and nonlinear transformations (Figure 2). To efficiently train such a DNN, Hinton and Salakhutdinov [49] introduced a greedy layer-wise unsupervised algorithm to initialize multiple-layer feedforward NNs. Since then, this training algorithm has been frequently shown to have a powerful capability to capture representative features via massive unlabeled data, and has obtained tremendous success in a variety of applications, particularly in the context of ASA [7], [50], [51]. The remainder of this section introduces several of the most important deep architectures for URL, including deep belief networks (DBNs), stacked autoencoders (SAEs), convolutional NNs (CNNs), and recurrent NNs (RNNs).

In contrast to data augmentation and speech synthesis, URL techniques attempt to leverage massive unlabeled data, rather than sparsely labeled data.

Deep belief networks and stacked autoencoders

Two of the most established deep-learning architectures are DBNs and SAEs. These topologies are formed by stacking multiple layers of restricted Boltzmann machines (RBMs) or feedforward autoencoders, respectively. The unsupervised pretraining of these architectures is done one layer at a time.

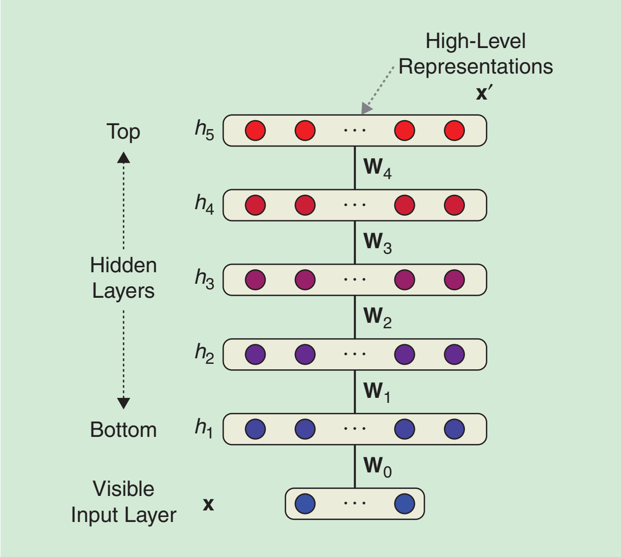

FIGURE 2. An illustration of typical deep URL. Usually, each layer of the network is individually trained in an unsupervised manner; this allows the network to incrementally learn a more robust representation than the one learned by training the network as a whole.

For SAEs, each layer is trained with an encoder $h(\cdot)$ and a decoder $g(\cdot)$ by minimizing the reconstruction error at its input $\mathbf{x}$:

$$g(h(\mathbf{x})) \approx \mathbf{x}. \tag{7}$$

The output of the encoder $h(\mathbf{x})$ forms an alternative representation of the input $\mathbf{x}$ and is fed into the successive layer as input. This procedure is repeated layer-by-layer until all predefined layers are initialized. The training of the stacked layers in this manner allows a deep network to incrementally learn a more robust representation when compared to training the whole network, in ensemble, from a random initialization of weights. For further insights into the advantages of pretraining with autoencoders and RBMs, see [52]. This observation is particularly true for stacked denoising autoencoders [53], extensions of SAEs where the initial input $\mathbf{x}$ is partially corrupted into another version $\tilde{\mathbf{x}}$ by means of stochastic mapping, i.e., $\tilde{\mathbf{x}} \sim q_d(\tilde{\mathbf{x}} \mid \mathbf{x})$. The robustness of the high-level representations formed using this technique is improved when compared to the aforementioned SAE [53].

An early attempt at applying deep-learning technologies to learn speech representations was proposed by Deng et al. [54], where the authors utilized DBNs and deep SAEs to compress (represent) speech directly from spectrograms. When compared with the traditional compression approach of vector quantization, this technique showed a much lower log-spectral distortion over the entire frequency range of wide-band speech. Expanding on the work of this article, DBNs have been extensively tested as an acoustic modeling paradigm for speech recognition and have shown encouraging performance in comparison with the conventional Gaussian mixture model (GMM) and HMM-based acoustic models for ASR [51]. For an overview of deep URL models for ASR and the corresponding performance gains, the reader is referred to both [6] and [50]. Inspired by these achievements, deep URL techniques have started to become the dominant approach in almost all areas of speech processing.

Convolutional neural network

Another deep architecture currently exciting great interest is the CNN [55], [56]. CNNs are a biologically inspired variant of the multilayer perception (MLP) originally developed for visual perception tasks [55]. Typically, they consist of one or more convolutional layers (often with a subsampling layer), followed by one or more fully connected layers.

CNNs are normally trained in a supervised manner. However, unsupervised training approaches are gaining in popularity. Inspired by the unsupervised learning algorithm for DBNs, Hau and Chen [57] constructed a deep architecture using a CNN trained in an unsupervised manner as an alternative building block to retrieve effective hierarchical speech representations. Specifically, the authors utilized an unsupervised predictive sparse decomposition algorithm to train the weights of the encoder and decoder [57].

Furthermore, a combination structure of CNNs and DBNs was proposed in [58], in which the authors constructed convolutional DBNs (CDBNs) with convolutional RBMs (CRBMs) as the building blocks. The CRBM is an extension of the conventional RBM to a convolutional setting. The weights between the hidden units and the visible units are shared among all locations in the hidden layers [59]. By leveraging a large amount of unlabeled data, the authors demonstrated that the learned hierarchical CDBN representations are competitive with conventional features (e.g., MFCCs) when evaluated across multiple audio classification tasks.

Long short-term memory recurrent neural networks

Unlike the aforementioned NN architectures, RNNs allow cyclical connections, which consequently endow the network with the capability of accessing previously processed information (i.e., context sensitivity). An advanced version of this paradigm, the long short-term memory (LSTM)–RNN [60], has recently attracted a large amount of attention. An LSTM unit contains one input, one output, and one forget gate to control the memory cell, which enable it to store and access information over a long temporal range. Therefore, the LSTM–RNN combination has a powerful capability for sequence learning.

In utilizing the advantages associated with LSTM–RNNs, Srivastava et al. [61] recently proposed and explored an unsupervised sequence-to-sequence learning paradigm where the LSTM–RNNs are constructed as an encoder–decoder. By doing this, the system efficiently learns the underlying representations of video sequences for future frame prediction or sequence reconstruction. This model has been further investigated by Chung et al. [36] for audio segment representations, where the authors demonstrated its effectiveness for spoken-term detection when compared with classic DTW. More recently, the gated recurrent unit has emerged as a computationally simpler alternative to the LSTM unit [62].

Overall, deep unsupervised learning paradigms have seemingly great potential for learning useful representations of large-scale unlabeled speech data. Nevertheless, in most cases, it is necessary to implement additional supervised training, such as fine-tuning, to ameliorate the system for a specific application [51], [63]; therefore, a small amount of labeled data is often additionally required to produce state-of-the-art performance.

Semisupervised learning

Unlike URL, which aims to distill representative features from unlabeled speech, SSL is designed to enhance recognition models. Given a seed set of labeled data, SSL exploits information from a large set of unlabeled data in an efficient manner with minimal intervention from human annotators. SSL methods are generally distinguished as being conducted in either an inductive or transductive manner [88]. The primary discrepancy between them lies in whether the distribution information of the unlabeled data is utilized for their own prediction.

Inductive approaches require the construction of a classification model *f* based on a priori knowledge of labeled data. The predictive model *f* is then used for predicting the unlabeled data, no matter whether they are presented in an online (afterward) or offline (beforehand) manner. Hence, inductive approaches are also known as a *supervised learning + additional unlabeled data* paradigm. Mathematically, this can be expressed as

$$\{(\mathbf{x}^l, y^l), l = 1, \ldots, n_l\} \mapsto f,\ f \mapsto \{y^u, u = 1, \ldots, n_u\}. \tag{8}$$

Once the automatically predicted annotations have been obtained from the unlabeled data set L^*_{ssl}, the labeled training set is updated, i.e., $L' = L \cup L^*_{\text{ssl}}$.

In contrast, transductive approaches do not need to prebuild a classification model *f* but instead perform predictions directly on the unlabeled data by exploiting the joint probability distributions of labeled and unlabeled data sets. In this technique, the unlabeled data set should be available beforehand. When new samples arrive, the transductive algorithms have to be rerun, which consequently increases the computational load. Hence, the transductive approaches are also referred to as the *unsupervised learning + additional labeled data* paradigm. That is,

$$\{(\mathbf{x}^l, y^l), l = 1, \ldots, n_l\} \cup \{\mathbf{x}^u, u = 1, \ldots, n_u\} \mapsto \{y^u, u = 1, \ldots, n_u\}. \tag{9}$$

Note that both the inductive and transductive approaches can be jointly deployed, as in transductive SVMs in which unlabeled data are also considered when determining the hyperplane [89].

The ASA literature is dominated by inductive SSL approaches. This is possibly due to inductive approaches being more flexible to the availability format of unlabeled data (i.e., online or offline). Among the inductive SSL approaches proposed, self-training (i.e., self-teaching) is arguably the most representative and has been widely and efficiently used for ASR [71], [72], emotion recognition [90], and speaker identification [73]. (In the context of ASR, SSL is often referred to as *unsupervised learning* or *unsupervised training*.)

A typical self-training paradigm is based on prediction uncertainty. That is, those samples $\{\mathbf{x}'^u_i\}$ recognized with high confidence *C* are picked up and combined into a selected subset *S*, and those $\{\mathbf{x}^u_j\}$ with low confidence remain in the unlabeled data set *U*:

$$\underset{\forall \mathbf{x}'^u \in S}{C(\mathbf{x}'^u)} \geq \underset{\forall \mathbf{x}^u \in U \backslash S}{C(\mathbf{x}^u)}. \tag{10}$$

The selected data set $\mathcal{S}$ (together with their pseudolabels) is then combined with the initial training set *L* to form a new data set $(L' = L \cup L^{*}_{\text{ssl}})$, which is sequentially employed to refine the previous model and retest the remaining unlabeled data. This process is repeated several times to incrementally upgrade the initial model.

Self-training is simple and can be easily applied to an existing model. However, it is open to the risk of error accumulation, which is introduced by the selection of misclassified data in early learning iterations. Commonly used techniques to mitigate such a detrimental effect include 1) using an additional development partition to determine the stopping point of learning, 2) using generalized expectation maximization to assign weights to the automatically labeled data based on the prediction confidence [74], and 3) retesting previously selected data for subsequent reevaluations and selections, such that the mislabeled data in previous iterations are possibly corrected in future iterations with an improved model [91].

Another commonly used inductive SSL paradigm in ASA is cotraining. Compared with self-training, cotraining attempts to exploit the mutual information between two learners (trained on different views or feature domains $\mathcal{X}_1$ and $\mathcal{X}_2$). That is, each learner uses its own predictions to teach not only itself, but also the other learner [92].

Successful cotraining relies on two assumptions: sufficiency and conditional independence [92]. Sufficiency infers that each view is sufficient for classification on its own, i.e., the two hypotheses $f_1\colon \mathcal{X}_1 \mapsto \mathcal{Y}$ and $f_2\colon \mathcal{X}_2 \mapsto \mathcal{Y}$ are good enough for recognition. Conditional independence denotes that the views are conditionally independent, given the class label, i.e., $P(y_i|\mathbf{x}) \leftarrow P(y_i|\mathbf{x}_1)P(y_i|\mathbf{x}_2)$. Although these two assumptions are restrictive, the work presented in [76] shows the capability of cotraining for retrieving emotional information in unlabeled data via separating the acoustic feature set into two pseudo views (i.e., not completely conditional independence) in the speech domain. Similar verification of cotraining has also been reported for other computational paralinguistics tasks [76]. Additionally, a more general framework called *multiview learning* requires less restriction in terms of conditional independence than cotraining and has been successfully applied in speech recognition by using several types of acoustic features and randomized decision trees [77].

More recently, SSL research in ASA has started to explore the advantages of deep-learning techniques [75], [93]. A typical implementation is ASR for a low-resource language [75], [93]. First, an initial DNN is trained in an unsupervised manner using multilingual data to learn the generalized representation of speech. Next, this model is fine-tuned as a seed model by using limited amounts of monolingual data from the low-resource language. The seed model is then employed to decode the untranscribed utterances, with the predicted hypotheses being regarded as the training transcripts for the next iteration. Various discriminative criteria (e.g., maximum mutual information or minimum cross entropy) can be adopted to obtain the prediction confidence scores for each frame, word, or utterance [75], [93]. Similar to traditional self-training and cotraining, the data (i.e., frame, word, or utterance) predicted with high confidence are assumed to be of high quality and are then incorporated to update the initial DNN or GMM–HMM acoustic model.

Apart from the inductive approaches, a graph-based transductive approach can also be integrated into DNN-based speech recognition systems at either a late or early stage [78]. For the late-stage integration, a graph is first constructed over the labeled and unlabeled data sets, where the node represents a data instance and the edge indicates the similarity between a data instance pair. Then, using a graphic-based learning algorithm, a new set of posterior distributions for each instance of unlabeled data is produced. After that, the posteriors are converted into a graph likelihood and are integrated with the original acoustic scores given by the DNN for a subsequent rescoring of the unlabeled data [78]. A major drawback of this late integration approach is a substantially increased computational cost, as the graph has to be reconstructed after each learning iteration. To overcome this problem, an early-stage integration algorithm has been proposed [78]. This algorithm employs a graph embedding approach in which the data in the graph is transformed into a compact feature vector, which is then used as additional input for the DNN.

Active learning

Similar to SSL, AL attempts to improve recognition models by exploring unlabeled data. However, unlike SSL, which performs automatic machine (model) annotation, the focus of AL approaches is to efficiently select the most informative data S in the unlabeled collection U for manual annotation. Partly because of the growing amounts of data to be handled and the popularity of crowdsourcing (see the "Efficient Data Labeling: Crowdsourcing" section), AL strategies for ASA are currently more important than ever.

One of the central goals of AL is to determine the informativeness of unlabeled data, a process known as *query strategy*. The following sections briefly review the most commonly used strategies with relevance to ASA, which include the uncertainty sampling, query by committee, and metaquery strategies.

Uncertainty sampling

This strategy uses confidence measures as a criterion to select the most informative data. The basic idea is to use a pretrained model (an active learner) to determine the uncertainty of predictions for a specific ASA task. The instances with the least certain predictions are then sent to an oracle (a human) for the annotation.

Formally, the selected data can be expressed as

$$\mathbf{x}' = \underset{\mathbf{x} \in U}{\arg\min}\, Q_c(\mathbf{x};\theta), \tag{11}$$

where θ indicates the model parameters trained on the labeled data set L and Q_c denotes the confidence measure function.

When using a probability model (e.g., Bayesian networks), this function is usually estimated using either the posterior probability, the probability margin between the two most likely class labels, or the entropy of prediction [94]. In the context of speech recognition, word posterior probabilities or the HMM-state entropy are frequently used as confidence measures [79], [81]. When using a nonprobability model (e.g., an SVM), similar measures can be constructed from discriminant functions. Considering the SVM as an example, pseudoprobabilistic values can be transformed from the output distances from the SVM hyperplane (see [17] for more details). The effectiveness of this approach has been extensively assessed for emotion recognition from speech [83].

Despite the reported performance improvement, many studies have found that uncertainty-based AL is inclined toward selecting noise and garbage data (i.e., outliers from the main data distribution) for human labeling. This issue occurs even more frequently when using AL to annotate data collected in the wild, i.e., not under controlled laboratory conditions, where environmental noises severely distort the speech, and many unexpected words are potentially uttered. Labeling these outliers is usually difficult and time consuming [95]. Furthermore, these data often offer little information on the overall system performance [17], [95]. A straightforward solution to address this outlier problem is to raise the threshold of a confidence score. For example, the authors of [17] used a median uncertainty strategy instead of the least certainty one for actively selecting spontaneously emotional utterances, which delivered a positive performance improvement.

Sampling by uncertainty and density (SUD) is a more sophisticated method that was introduced for ASR in [96]. In this approach, unlabeled instances that are both near the decision boundary and very close to other examples are assumed to be more important than those that are isolated (i.e., likely to be outliers). Hence, SUD considers not only the most informative data in terms of uncertainty but also the most representative data in terms of density. That is, those data predicted with least certainty and distributed in a low-density area are ignored.

A similar idea was proposed in [80], where the global criterion was used in ASR to maximize the expected lattice entropy reduction over all nontranscribed data. Specifically, it first measures the entropy among the lattices generated by decoding unlabeled utterances. It then estimates the expected entropy reduction over the whole data set for each given utterance, and selects the utterances that should deliver the highest entropy reduction for human labeling. After that, the transcribed utterances can be weighted according to the number of similar utterances in the whole data set to achieve better performance for speech recognition. This algorithm is also analogous to the error-rate reduction strategy introduced in [95].

Query by committee

This strategy uses a committee (group) of weak models (learners), denoted by $\Theta = \{\theta_1, \ldots, \theta_k\}$, to select unlabeled data by the principle of maximal disagreement among these models [97]. Mathematically, this can be expressed as:

$$\mathbf{x}' = \underset{\mathbf{x} \in U}{\arg\max}\, Q_d(\mathbf{x}; \Theta). \tag{12}$$

The two key problems in committee-based approaches are 1) constructing a committee Θ that represents competing hypotheses and 2) defining a disagreement measurement Q_d. To alleviate the first problem, the models are usually built by employing multiple different classifiers (e.g., HMMs, SVMs, and RNNs) with the same training data, or by splitting the training data or features into partitions for training several different versions of the same type of classifier, or by a combination thereof. For the second problem, the commonly used disagreement measures are vote entropy and Kullback–Leibler divergence (see [94] for more details). In speech recognition, this strategy has been applied to both acoustic and language models, resulting in a significant data annotation reduction while achieving the same word accuracy [82].

Meta query strategies

One often deals with imbalance across classes of interest in the data. As an example, for emotion recognition, the emotional speech of interest usually appears sparsely within a data set, while the less interesting nonemotional speech often appears at a much higher frequency. In this scenario, an initial coarse model can be used to first decide which data are of interest by distinguishing between neutral and emotional speech. A subsequent finer model can be then used to recognize different emotions or respective other classes in other tasks in the selected emotional speech data. An example of such an approach is the sparse-tracking query strategy [83]. It tracks only sparse (emotional) instances, via iterative retraining and labeling, using a novelty detection paradigm.

One issue when analyzing subjective speaker states and traits (e.g., emotion and personality) is the requirement of multiple annotations per sample to obtain a reliable gold standard, which linearly increases the annotation workload. Recently, dynamic active query strategies have been shown to be successful in overcoming this issue [84]. These approaches, e.g., sequentially query human annotators to label a specific instance up to the achievement of a predefined agreement level (i.e., a certain number of votes for a specific class). The general idea is to learn and exploit the varying reliability of raters to discern whom to best trust and when. The results presented indicate that this approach can contribute to a meaningful reduction of annotation effort [84].

Cooperative learning

As discussed previously, SSL techniques can perform annotation work from machines with a bare minimum of human intervention. However, the performance of SSL is hampered by the issue of potential error accumulation [94]. Alternatively, AL techniques have the potential to achieve higher accuracy with fewer training labels by actively selecting the data it can learn the most from. However, AL still requires a considerable amount of human intervention.

To take advantage of the best of both approaches, it is plausible to jointly conduct AL and SSL in a unified CL framework [17]. A general CL flowchart is illustrated in Figure 3. CL allows the sharing of the labeling effort between human and machine oracles, while being able to mitigate the limitations of SSL and AL. This is achieved by successively fusing the data subset selected by the AL (L_{al}) and the one selected by SSL (L^*_{ssl}) into the original training set in an iterative fashion. In this case, the labeled data set L' is continuously updated by $L' = L \cup L_{\mathrm{al}} \cup L^*_{\mathrm{ssl}}$. To minimize the effects relating to error accumulation, AL is often conducted before SSL.

Early studies of CL mainly focused on text classification. McCallum and Nigam were the first to investigate the idea of integrating the query by committee-based AL and the expectation maximization-based SSL for text classification [98]. Later, motivated by the success of cotraining (see the "Semisupervised Learning" section), a similar idea of jointly using multiple views was taken into account, contributing to the new CL algorithm of coexpectation-maximization testing [99].

For speech processing, the first CL efforts were undertaken by Riccardi and Hakkani-Tür [85] for ASR. This approach assigned confidence scores to transcribed utterances based on the lattice output, from which the utterances were determined to be manually or automatically labeled. A similar idea was also investigated by Yu et al. [86] for speech recognition. In this approach, the data recognized with high confidence are translated automatically by machine, while the ones recognized at a low confidence are selected and translated manually. Similar to the uncertainty-based AL, this uncertainty-based CL is as well inclined to choose noise and garbage utterances that typically have low confidence scores.

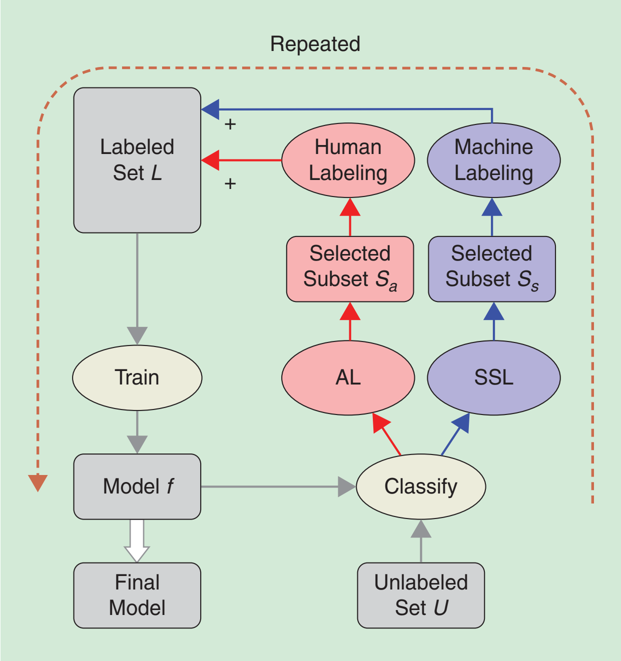

FIGURE 3. A general overview of a CL framework that aims to take advantage of both AL and SSL.

Motivated by the success of the global entropy reduction maximization criterion [80] for AL (see the "Active Learning" section), Yu et al. [87] extended the work of [80] by integrating this approach with SSL. The results presented indicate that this technique achieves a notable performance increase when compared to the uncertainty-based CL approaches for speech recognition. Besides, Zhang et al. [17] recently combined SSL with a median uncertainty-based AL for emotion recognition, which efficiently helps to avoid choosing garbage data as well. Furthermore, in the same article, multiview CL (i.e., where two views are used for both AL and SSL) was exemplified and demonstrated to achieve better performance than the single-view CL [17].

CL is indeed a productive, highly efficient way to exploit unlabeled speech data to enhance the performance of preexisting models while minimizing human work.

Experimental results obtained in the aforementioned studies indicate that, when compared to SSL and AL, CL is indeed a productive, highly efficient way to exploit unlabeled speech data to enhance the performance of preexisting models while minimizing human work. Moreover, its potential is expected to be further evoked when implemented with a crowdsourcing platform (see the "Efficient Data Labeling: Crowdsourcing" section and/or, incorporated with deep-learning techniques, the "Unsupervised Representation Learning" section).

Learning from unreliable or unbalanced resources

In contrast to both the no- and limited-resource techniques, which address the speech data quantity challenge, this section focuses on the methods that aim to tackle the speech data quality challenge. In particular, it covers techniques designed to operate in the presence of unreliable or unbalanced resources.

Data selection

Data quantity and diversity are both vitally important properties when building a robust ASA system. However, they can introduce a range of confounding factors. For example, speech utterances that are severely distorted by noise might be present in a prototypical data set. Owing in part to a lack of annotators' concentration, these data are often improperly labeled or even mislabeled. This gives rise to the necessity of data selection to discard such garbage data, as accurate decisions made by a pattern recognition engine are largely related to high-quality training data.

The goal of data selection is to select a smaller data source S that is most representative (i.e., most informative) of the entire data L, i.e., $S = DS(L)$ and $S \subseteq L$, thus omitting any superfluous or garbage data. The concept of data selection discussed in this section differs from that for AL or SSL, which is carried out on unlabeled data (see the "Semisupervised Learning" and "Active Learning" sections). It also differs from feature selection methods (e.g., filter or wrapper selection), which select the most informative features for a particular ASA task. Instead, the data selection techniques reviewed are designed to select labeled samples or instances that will serve as learning units.

Within the ASA literature, Wu et al. [100] selected the samples that had a uniform distribution across speech units (i.e., words and phonemes) by the principle of maximum entropy for ASR. The experimental results presented indicate that a system trained on a 150-h selection of data could achieve competitive results with a system trained on the full 840-h data set.

When performing subjective ASA recognition tasks (e.g., emotion recognition), a learning and testing target has to be generated usually by fusing the labels of multiple annotators to reduce subjectivity. In addressing the unreliable label problem, Erdem et al. [101] performed the RANSAC data selection algorithm to remove potentially mislabeled instances when training a model, and obtained better emotion recognition performance. This algorithm operates in an iterative fashion. First, it uses a small subset of the data to determine the initial model parameters. Then, the unused data instances are tested against this model, and those that fit the model within a predefined tolerance, denoted as ϵ, are considered to be a part of the consensus set. When the size reaches a predefined limit, the model parameters are updated using all of the consensus data and initial data. This procedure is repeated several times. More recently, Zhang et al. [102] reported that annotation reliability can be assessed using the human-agreement level among multiple annotators. Data with a low human-agreement level are considered to be mislabeled data and are removed from the data set.

TL approaches can be mostly grouped into one of three categories according to the properties of the knowledge transferred: instance-, feature-, and model-based TL.

Data balancing

When collecting data for a specific ASA task, such as modeling speaker states (e.g., affection or intoxication) or characteristics (e.g., likeability), one often faces issues relating to class scarcity. While interesting speech samples are required, the majority of the ubiquitous speech data are essentially neutral. This can result in highly imbalanced class distributions and recognition systems that perform poorly when attempting to recognize the target classes [103].

Numerous studies in the context of machine learning have tackled this issue by data balancing [103], with the purpose of balancing the data distribution over classes, i.e., $L_{\mathrm{bl}} = L_1 \cup L_2 \ldots \cup L_n$ where $L_1, L_2, \ldots, L_n$ denote labeled data from n different classes that contain approximately the same amount of data. Among the methods proposed, data sampling is seen as a simple and efficient method. Data sampling is the process of either repeating preexisting data, regenerating new data to modify the imbalanced data distribution, or randomly removing part of the data to produce a data set with a more balanced class distribution.

One common method is random sampling, either by oversampling (i.e., upsampling) or by undersampling (i.e., downsampling). The former approach essentially involves randomly selecting a subset of instances $L'_{\min}$ in the minority class $L_{\min}$ and adding them back into the original training set L, $L = L \cup L'_{\min}$. In contrast, the latter technique involves the random selection of a subset of instances L'_{maj} in the majority class L_{maj} and removing them from the original training set L, $L = L \setminus L'_{\mathrm{maj}}$. However, this process may result in a loss of important information pertaining to the majority class.

Another frequently used and effective method for data sampling is SMOTE [104]. The underlying idea is the creation of a new set of artificial examples belonging to the minority class. Data sampling has been widely used for computational paralinguistics with notable effects [17], [105]. Even in ASR systems, balancing the sample distributions among all phonemes has been shown to outperform the baseline by a large margin [106].

Learning from unmatched resources

Conventional machine-learning approaches operate under the assumption that instances from both the source and the target domains are independent and identically distributed. However, in real-world scenarios, this is very rarely the case; one will inevitably encounter the problem of distribution mismatch (also known as the *data set bias*) or covariate shift between the data in the target and source domains (i.e., $\mathcal{S} \neq \mathcal{T}$). Such discrepancies often give rise to a substantial downgrade in the performance of affected speech analysis systems. TL is a potential solution to bridging the mismatch gap.

The objective of TL is to improve the predictive function in the target domain $\mathcal{T}$ using the knowledge from a different but related source domain $\mathcal{S}$ (Figure 4). A wide range of TL approaches have been proposed in the machine-learning and data-mining literature. TL has also been applied to many ASA tasks, including low-resource language ASR, speaker adaptation, and emotion recognition.

TL approaches can be mostly grouped into one of three categories according to the properties of the knowledge transferred: instance-, feature-, and model-based TL. These

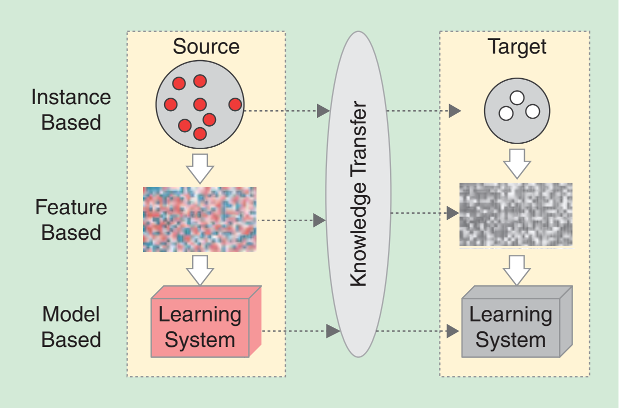

FIGURE 4. An illustration of TL: knowledge learned in the source domain is used to aid analysis in the target domain. This transfer can take place at either the instance, feature, or model level.

Table 2. Selected TL studies on the unmatched speech resource.

Publications	Types	Approaches	Models	Applications	Databases and Languages
Hassan et al. 2013 [111]	Instance	KMM, KLIEP, uLSIF	SVM	ER	FAU AEC (Ge)
Doulaty et al. 2015 [113]	Instance	Submodular data selection	DNN	ASR	Data collected in six settings
Narayanan and Wang 2013 [115]	Feature	Denoising	DNN	ASR	Aurora-4
Deng et al. 2013 [116]	Feature	SAE	SVM	ER	Six emotional corpora
Kocscor and Tóth 2004 [117]	Feature	KPCA, KLDA	GMM, ANN, etc.	Vowels/phoneme classification	Hungarian (Hu), TIMIT (En)
Jafari and Plumbley 2011 [118]	Feature	Sparse coding	/	Speech representation/ denoising	Freesound
Dahl et al. 2012 [51]	Feature	DNN, signal task	DNN–HMM	ASR	Bing mobile voice (En)
Amodei et al. 2015 [7]	Feature	CNN, signal task	CTC–RNN	ASR	English (En) and Mandarin (Ma)
Heigold et al. 2013 [119]	Feature	SHL–DNN, multitask	Softmax layer	Multi-/cross lingual ASR	Data in various languages
Huang et al. 2013 [120]	Feature	SHL–DNN, multitask	Softmax layer	Multi-/cross lingual ASR	English (En) and Mandarin (Ma)
Miao et al. 2015 [121]	Feature	SAT–DNN, *i*-vector	DNN	ASR	TEDLIUM (En)
Deng et al. 2014 [122]	Feature	SHL–DNN	SVM	ER	Three emotional corpora
Giri et al. 2015 [123]	Feature	SHL–DNN	DNN	Robust ASR	REVERB Challenge corpus (En)
Leggetter and Woodland 1995 [124]	Model	MLLR	GMM–HMM	ASR	ARPA RM (En)
Deng et al. 2014 [112]	Model	DAE, multitask	SVM	ER	Three emotional corpora

En/Ge/Hu/Ma: English/German/Hungarian/Mandarin; ER: emotion recognition; uISIF: unconstrined least-squares importance fiting; KLDA: kernal linear discriminant analysis; SHL: shared hidden layer; SAT: speaker adaptation training; DAE: denoising autoencoder; KPCA: kernel principal components analysis.

approaches as well as data agglomeration are elaborated upon in the following. These sections are intended to be a succinct overview of these techniques for ASA. For a more general survey of TL, see [20] and [21]. A selection of typical TL studies for ASA are listed in Table 2.

Instance-based TL

Instance-based TL assumes that certain subsets of the data in the source domain can be used for learning in the target domain by means of reweighting. Instance-based TL essentially assigns more weight to those source domain data that are similar in terms of distribution to the target data, and less weight to those that poorly reflect the distribution of the target data. The technique of weighting the input data based on the target data is known as *importance weighting* for covariate shift or sample selection bias. With the aim of minimizing the expected classification error, the estimation of the importance weights β is achieved as a ratio calculation problem:

$$\beta(\mathbf{x}) = \frac{P_{\mathcal{S}}(\mathbf{x})}{P_{\mathcal{T}}(\mathbf{x})}, \tag{13}$$

where $P_{\mathcal{S}}(\mathbf{x})$ and $P_{\mathcal{T}}(\mathbf{x})$ are the probability densities of the source and target domain data, respectively [107].

Instance-based TL assumes that certain subsets of the data in the source domain can be used for learning in the target domain by means of reweighting.

The most straightforward approach to calculating this density ratio is to directly estimate the target and source densities separately. However, this approach tends to perform poorly because of the inherent difficulty of density estimation, particularly in high-dimensional cases. In this regard, instance-based TL techniques, which estimate the importance ratio without estimating the densities, have been proposed. For example, Huang et al. [108] proposed a kernel-based method known as *kernel mean matching* (*KMM*). It reweights the instances by matching the means between the source domain data and the target domain data in a reproducing-kernel Hilbert space. The downside of KMM is that its performance is highly dependent on the choice of hyperparameters (model selection), which need to be heuristically tuned.

To overcome this issue, Sugiyama et al. [109] introduced the Kullback–Leibler importance estimation procedure (KLIEP) algorithm. KLIEP estimates the importance ratio by minimizing the Kullback–Leibler divergence between the original target data density and its corresponding estimation. Owing to the convex property of the involved optimization problem, the KLIEP algorithm can obtain unique global solutions. In addition, the tuning parameters can be objectively optimized, based on a variant of cross validation. While KLIEP is seemingly more advantageous than KMM, it is actually less computationally efficient because of the high linearity of the objective functions to be optimized.

This issue was addressed by Kanamori et al. [110] by means of least-squares importance fitting (LSIF). The LSIF algorithm formulates the direct importance estimation problem as a least-square function fitting problem: casting the optimization problem as a convex quadratic program that can be efficiently solved using a standard quadratic program solver. This algorithm was further extended to be unconstrained LSIF (uLSIF), which greatly improved the computational efficiency [110]. For emotion recognition, the approaches of KMM, KLIEP, and uLSIF have shown great success in alleviating the discrepancy between different speech resources [111], [112].

An alternative to the aforementioned approaches is binary reweighting. It selects the data from the source domain based on the data distribution to reduce the discrepancy between the source domain and the target domain. This strategy is related to the data selection strategy used for AL (see the "Active Learning" section), which can be viewed as a specific data selection case in a source-data unlabeled setting. It is also related to the data selection strategy discussed in the "Data Selection" section, which attempts to improve the quality of the data only in the target domain.

A prominent binary reweighting approach is based on using submodular functions to simulate the acoustic similarity between the target and source domain data.

A prominent binary reweighting approach is based on using submodular functions to simulate the acoustic similarity between the target and source domain data [113], [114]. The process identifies a subset L' of the complete source data set $L_{\mathcal{S}}$, so that any subsequent subset L'' added to this selected subset will not increase the value of the submodular function f, i.e., $L' = \arg\max \{f(L' \cup L'') < f(L'), \text{ where } L' \subseteq L, L'' \subseteq L \setminus L'\}$. In doing this, only the positive transfer is exploited across domains. In ASA, submodular function-based data selection has been extensively evaluated for multidomain speech recognition and has shown superior performance [113], [114].

Feature-based transfer learning

The goal of feature-based TL approaches is to find a transformation function $\Phi(\cdot)$ that can be used to convert the source feature space and/or target feature space into an approximately matched distribution space while preserving the important properties of the original data. Mathematically, this can be expressed as

$$P(\Phi_{\mathcal{T}}(X_{\mathcal{T}})) \approx P(\Phi_{\mathcal{S}}(X_{\mathcal{S}})), \tag{14}$$

or

$$P(Y_{\mathcal{T}} | \Phi_{\mathcal{T}}(X_{\mathcal{T}})) \approx P(Y_{\mathcal{S}} | \Phi_{\mathcal{S}}(X_{\mathcal{S}})). \tag{15}$$

In achieving this, two possible strategies exist: asymmetric and symmetric strategies. The asymmetric strategy keeps either the source or target feature space unchanged, and maps the other one onto it (i.e., $\Phi_{\mathcal{T}} : \mathcal{T} \rightarrow \mathcal{S}$ or $\Phi_{\mathcal{S}} : \mathcal{S} \rightarrow \mathcal{T}$). By contrast, the symmetric strategy transforms both source and target feature spaces into a new latent one (i.e., $\Phi_{\mathcal{T}} : \mathcal{T} \rightarrow \mathcal{Z}$ and $\Phi_{\mathcal{S}} : \mathcal{S} \rightarrow \mathcal{Z}$), in which they share the same distribution and knowledge relationship.

In achieving this, two possible strategies exist: asymmetric and symmetric strategies. The process of denoising distorted (noisy) speech can make the feature space (target) of noisy speech closer to that of clean speech (source). In doing this, the cleaned speech can be evaluated by preexisting acoustic models, which are often trained on the clean speech. An emerging research trend in the speech enhancement community is to use DNNs (e.g., deep LSTM–RNNs) to map noisy speech into its clean counterpart or ratio mask on a frame-by-frame basis. Preliminary results have proved that this method is quite effective, particularly for alleviating nonstationary noise [115]. For more details of speech denoising technologies, see [125].

Apart from speech denoising, a more general TL method to reduce the database bias was proposed in [116] and is based on an SAE—an autoencoder with sparsity enforced in the hidden layer (see the "Unsupervised Representation Learning" section). This method is a fully supervised approach. First, using the target data, class-specific SAEs are trained, and then treated as the transforming models ($\Phi(\cdot)$). The source data are then fed into SAEs corresponding to its class, and thus a new source representation is constructed. In doing this, the distribution of the new source feature space is expected to be inclined to the target one. Finally, the new source data are used to train a standard classifier.

As for the symmetric strategy, early studies were mainly conducted using principal component analysis (PCA), linear discriminant analysis (LDA), and sparse coding. The goal of these approaches is to learn a low-dimensional latent feature space or a shared space. The resulting feature space can serve as a bridge for transferring meaningful knowledge from the source domain to the target domain [20]. PCA is typically used to project the data along the direction of maximal variance in an unsupervised way. LDA, or Fisher's LDA (FDA), on the other hand, is used to project the data onto a line that can maximize the distance between the means of the two classes (in a binary classification case) while minimizing the variance within each class.

Both PCA and LDA are linear transformations that limit their applicability to most real-world data. In this regard, kernel functions (e.g., Gaussian, Cauchy, and polynomial kernels) can be used in conjunction with PCA and FDA, resulting in kernal PCA (KPCA) and kernel FDA (KFDA) paradigms that transform data in a nonlinear manner. Owing to their simplicity and effectiveness, KPCA and KFDA have been widely used in the speech processing community [117]. Similarly, kernel canonical correlation analysis has been applied to cross lingual emotion recognition [126].

Sparse coding, also termed *dictionary learning*, attempts to find succinct representations (i.e., atoms or elements of the dictionary) of the input data such that the input data can be represented as a linear combination of these sparse representations [127]. Compared to the aforementioned feature transformation methods, sparse coding has been demonstrated to be able to produce a more robust signal representation in speech reconstruction and denoising tasks [118].

Conventional feature transformation approaches are typically executed at a shallow level. Recently, deep-learning approaches for feature-based TL have begun to attract a lot of of research attention. Deep learning is regarded as a natural TL paradigm; it provides a powerful capability of learning high-level abstracts or representations that are more robust against the variation of conventional speech features (i.e., log Mel-filter banks and MFCCs) over different domains [50] (see the "Unsupervised Representation Learning" section). These representative features can then be used as normal features to train conventional discriminative or generative models, such as NNs, HMMs, and SVMs. Thanks to the invariant property of these representations, they can potentially deliver remarkable performance improvements for almost all ASA tasks [7], [50], [51], [58].

Researchers have started to investigate the learning of robust representations over multiple modalities (e.g., audio and video).

In addition to the basic representation learning approaches mentioned previously, more advanced topologies have begun to emerge, which explicitly involve several related tasks in a multitask learning paradigm. Multitask learning is the process of learning multiple tasks at the same time to learn a shared representation among different tasks. Mathematically, when training the model with multiple tasks, we aim to minimize the objective function as follows:

$$\mathcal{J}(\theta_0) = \sum_{k=1}^{K}\sum_{i} L(\mathbf{x}_{ki}, y_{ki}; \theta_k) + \frac{\lambda}{2}\|\theta_0\|^2, \tag{16}$$

where K is the number of tasks, $L(\cdot)$ denotes the loss function, and θ_0 stands for the general model parameters.

When performing deep multitask learning for multilingual or cross lingual speech recognition, it is typical to share the hidden layers across all languages [119], [120]. If learned appropriately, the hidden layers serve as increasingly complex feature transformations, sharing common hidden factors across the acoustic data from different languages. The final softmax layers, however, are not shared. Instead, each language has its own softmax layer to estimate the posterior probabilities specific to that language, using the most abstract representation from the topmost hidden layer. The strong result gained using this topology [119], [120] indicates its potential; it opens up the possibility for quickly building a high-performance recognition system for a new language using an existing multilingual DNN.

Many other deep multitask learning derivatives have been investigated to overcome the feature variation problems caused by factors such as different speaker characteristics, noisy environments, and poor recording channels. For example, Deng et al. [122] treated different corpora as different tasks for emotion recognition; Giris et al. [123] regarded noise type as an auxiliary task for speech recognition; and Seltzer and Droppo [128] treated phone label, phone text, and state context as different tasks when performing phoneme recognition. Recently, a universum autoencoder was proposed [129]. This technique uses a small amount of labeled data from the target domain and unlabeled data from a source domain to jointly minimize the reconstruction error and the universum leaning loss. Motivated by these achievements of learning representations among multiple related tasks, researchers have started to investigate the learning of robust representations over multiple modalities (e.g., audio and video) [130]. This topic, however, is beyond the scope of this overview.

Model-based transfer learning

Model-based TL, also known as *parameter-based TL*, aims to learn a new model from an existing model that has been well trained on rich source data. Unlike feature-based TL approaches, which usually transform the feature spaces, model-based TL approaches modify the pretrained model parameters (θ) to account for the differences that may exist between the domains. This can be formulated as

$$P(X_{\mathcal{S}}, Y_{\mathcal{S}}; \theta_{\mathcal{S}}) \rightarrow P(X_{\mathcal{T}}, Y_{\mathcal{T}}; \theta_{\mathcal{T}}) \tag{17}$$

for a generative model or

$$P(Y_{\mathcal{S}} | X_{\mathcal{S}}; \theta_{\mathcal{S}}) \rightarrow P(Y_{\mathcal{T}} | X_{\mathcal{T}}; \theta_{\mathcal{T}}) \tag{18}$$

for a discriminative model.

Early-stage model-based TL approaches in the speech community included maximum a posteriori (MAP) estimation and maximum likelihood linear regression (MLLR), which are designed for generative models (e.g., GMM–HMM). These techniques have been applied successively to speaker adaptation [131], where the speech from each specific speaker is supposed to be in a different domain with the initial training data. They have also been shown to be useful in computational paralinguistics tasks, such as depression detection [132].

Specifically, MAP uses the speaker-independent models (i.e., universal background models) as a prior probability distribution over the model parameters, and then performs maximum likelihood estimates by considering the model parameters obtained on the speaker-dependent data. Alternatively, MLLR calculates a set of linear regression transformations to shift both the means and the covariances in an initial Gaussian mixture HMM system so that each state in the system is more likely to have generated the speaker data the model is being adapted to [131]. Compared with MAP, MLLR requires fewer adaptive data. Aside from speaker adaptation, these methods have been applied to other acoustic variation adoption scenarios, such as noise adaptation [125].

Due largely to the recent advancements in deep learning, discriminative model-based TL has recently become an active research topic. In deep learning, the simplest way to adjust the pretrained model parameters when adapting to a specific task is through fine-tuning. As discussed in the "Unsupervised Representation Learning" section, pretraining is a down–up unsupervised algorithm, which can be considered as a model initialization process that attempts to produce a model that has a global optimization attribute. By contrast, fine-tuning is an up–down supervised algorithm to optimize all of the NN weights jointly with the labeled target data. This procedure is usually performed using backpropagation of error derivatives [63].

In contrast to the more sophisticated TL approaches discussed, a simpler solution to utilize multiple sources of data is data agglomeration.

Another paradigm to adapt the model to the target data, the adaptive denoising autoencoder, is highly related to multitask learning [112], [133]. This paradigm is usually undertaken in two steps. In the first step, a source model is trained on the source data. In the second step, the trained model parameters are used as prior information to regularize the adaptation process of the model on the target data, so as to minimize the objective function as follows:

$$\mathcal{J}(\theta_{\mathcal{T}}) = \sum_{i=1}^{n_{\mathcal{T}}} L(\mathbf{x}_i, y_i; \theta_{\mathcal{T}}) + \frac{\lambda}{2}\| \theta_{\mathcal{T}} - \beta\theta_{\mathcal{S}} \|^2, \qquad (19)$$

where $n_{\mathcal{T}}$ is the number of labeled target data, $L(\cdot)$ denotes the loss function on the target data, $\theta_{\mathcal{S}}$ represents the well-trained model on the source data (source model), $\theta_{\mathcal{T}}$ denotes the expected new model on the target data (target model), and β is the adaptation coefficient. Since the discrepancy between the source and target models is explicitly considered as a penalty term in the objective function, this approach is also known as *regularized adaptation* [133]. In emotion recognition applications, this approach has started to show promising results [112]. Note that such model-based multitask learning paradigms differ from the feature-based approaches covered in the "Learning from Unmatched Resources" section, where the model is trained in only one step by calculating the joint loss of all of the tasks in the objective function [see (16)].

Data agglomeration

In contrast to the more sophisticated TL approaches discussed, a simpler solution to utilize multiple sources of data is data agglomeration [134]. In this approach, one or more source databases are directly concatenated with the target database to form a large-size data pool $P = L_{\mathcal{T}} \cup L_{\mathcal{S}_1} \cup \ldots \cup L_{\mathcal{S}_k}$. This approach is suitable only when the various data sources are for similar tasks and share a common feature set.

To help ease any potential database biases, it is desirable to apply 1) normalization techniques such that the scattered feature spaces can be unified into a shared one and 2) task mapping to retain label consistency. The three normalization methods frequently applied in the literature are centering, min–max normalization, and standardization. Applied not only to each corpus separately (i.e., before data agglomeration), these methods can be also used after building a joint training set from multiple databases. Thanks to these normalization approaches, data agglomeration has been frequently applied to, e.g., emotion recognition [134]. As for task mapping, it is necessary to find the relationship between different tasks. For example, in emotion recognition, the prototypical emotions (e.g., anger, contempt, disgust, fear, interest, joy, sadness, and surprise) can be mapped onto the emotional dimensions of arousal and valence [134].

Conclusions and challenges for future work

To continue building on the success of machine-learning methods for ASA, there is a need for large amounts of labeled data. However, the work of collecting such data is costly and time consuming. Clever engineering can go a long way toward solving this problem by helping to leverage unlabeled, unreliable, or unmatched data. Motivated by this, we systematically presented an overview of the very recent and prominent techniques that intend to semiautonomously enrich the data quantity and enhance the data quality.

Crowdsourcing was discussed as an efficient data annotation approach, with the caveat that it requires quality control management. The integration of crowdsourcing with AL or CL strategies to intelligently and dynamically select data for labeling has the potential to further reduce the annotation workload and improve overall data quality.

Spoken-term detection and discovery and related means of retrieval of speech-related phenomena were discussed in relation to addressing the sparse data challenge. While these techniques can automatically find patterns in speech utterances without any labeled resource, the associated computational complexity limits their application to smaller databases. Reducing the computing complexity of these techniques is an essential direction of future research. Other techniques discussed on the sparse data challenge were data augmentation and speech synthesis. These techniques can artificially generate labeled speech data in a limited-labeled-resource setting. A key concern about their ongoing use is how to guarantee that the speech samples generated have a positive effect on the analysis being performed. Research into identifying task-invariant features has been identified as one potential solution in this regard.

With its capability to leverage information from large-scale unlabeled data, deep URL has delivered breakthrough results in a variety of ASA tasks. Future research efforts, particularly those focused on network construction strategies, are expected to increase the generalizability of the extracted features and thus improve on the already impressive capabilities of this paradigm. AL, SSL, and CL are other efficient techniques to take advantage of unlabeled data. In this regard, we identified the integration of SSL and deep learning as a particularly promising future research direction.

To handle the unreliable-data challenge, data selection and data-balancing techniques were also reviewed. Despite the conventionality of the reviewed algorithms, dynamically selecting and balancing data is of great importance to the machine-learning process. The role and importance of these well-practiced techniques in relation to deep learning are still being established.

To deal with the unmatched data challenge, TL strategies and data agglomeration were discussed. TL in particular, owing to its effectiveness, has attracted increasing amounts of research attention. However, when improperly used, these techniques substantially degraded overall system performance. Therefore, how to achieve positive transfer while preventing negative transfer between appropriately related tasks is an important and open research issue.

Although great opportunities are offered by the techniques reviewed, many additional risks may be brought to light through their practical application. For example, with the growing popularity of the use of microphones, the Internet, crowdsourcing, and cloud computing, personal speech signals easily run the risk of being disclosed to the public domain. Furthermore, from such data it is largely possible to extract confidential speaker information, such as a speaker's age, gender, or identity. Therefore, how to best protect the security and privacy of users has become a major area of concern in this field [135].

A potential solution in this regard is a distributed recognition system, such as the one proposed for computational paralinguistics in [136]. In this system, functionals are applied over the LLDs to extract features. These statistical features, rather than the LLDs or the raw signals, are transmitted from the client side to the server side. The procedure of generating these feature vectors is irreversible. Therefore, as the LLDs cannot be reconstructed, the contents of the original speech signals are protected. Recently, a decentralized SSL paradigm was proposed in [137], in which privacy-preserving matrix completion algorithms are used, so that only learned knowledge is transferred between different clients, while the raw data are incommutable. However, as these approaches cannot fully guarantee client security and privacy or maintain the original performance, continued research addressing privacy concerns is required.

The techniques discussed in this article are mainly applied in an offline manner. However, the realistic application of a specific task offers the opportunity to collect truly massive amounts of real-world data in an online fashion. For example, Google reported that 55% of teenagers and 41% of adults in the United States [138] used their voice search more than once a day in 2014. Hence, research is needed into techniques to dynamically make use of future data to enhance the adaptiveness of preexisting models to various speakers, environments, and tasks. Such techniques are commonly referred to as *online* and *incremental* learning [139], [140].

Finally, the recent developments in dialog management systems, the computerized spoken language understanding and generation of natural and meaningful responses during speech-based human–computer interactions, means it is now more feasible than ever to explore cues extracted from an entire conversation process to aid ASA systems. Such cues could indicate the correctness of previously performed analyses and as such would be considered a form of reward or punishment information. This information could be sequentially exploited using reinforcement learning strategies to dynamically update the decision mechanism of the predictive model. Deep reinforcement learning, in particular, has become an active and growing research topic in machine learning [141]. But despite being widely applied in related fields, such as dialog management, research into reinforcement learning for ASA is currently in its infancy. We firmly believe that research into deep reinforcement learning has the potential to move ASA technologies out of controlled laboratory settings and into diverse, practical everyday environments leading to more intelligent (even emotionally and socially intelligent) and adaptive ASA systems.

Despite these risks and challenges, the techniques reviewed in this article will play a key role in opening up new research opportunities to explore the value of big unlabeled, unreliable, and unmatched speech data. It is our strong belief that the continued growth in the research and applications discussed will facilitate the emergence of novel techniques to fill the gap between no-labeled-resource and reliable big data and usher in the next generation of ASA technologies.

Acknowledgments

This work was supported by the European Union's Seventh Framework Program through ERC Starting Grant 338164 (iHEARu), and by the Horizon 2020 Program through Research Innovation Action 688835 (DE-ENIGMA).

Authors

Zixing Zhang (zixing.zhang@uni-passau.de) received his B.S. degree from the Chinese Agricultural University in 2007, his M.S. degree from the Beijing University of Posts and Telecommunications, China, in 2010, and his Ph.D. degree from the Technische Universität München, Germany, in 2015. Currently, he is a postdoctoral researcher at the University of Passau, Germany. His research interests lie mainly in deep, semisupervised, active, and multitask learning; in the applications of human state and trait analysis from speech; and in robust automatic speech recognition. He is a Member of the IEEE.

Nicholas Cummins (nicholas.cummins@uni-passau.de) received his B.Eng. degree (first-class honors) in electrical engineering from the University of New South Wales (UNSW), Sydney, Australia and his Ph.D. degree in electrical engineering from UNSW in February 2016. His Ph.D. dissertation investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression. He is currently a postdoctoral researcher at the Chair of Complex and Intelligent Systems, University of Passau, Germany. His research interests include affective and behavioral computing. He is a Member of the IEEE.

Björn Schuller (schuller@ieee.org) received his diploma, doctoral degree, and habilitation degree in electrical engineering and information technology from the Technische Universität München, Germany in 1999, 2006, and 2012, respectively. He is a reader in machine learning in the Department of Computing at Imperial College, London, United Kingdom, and a full professor and head of the Chair of Complex and Intelligent Systems, University of Passau, Germany, where he previously headed the Chair of Sensor Systems. He is a Senior Member of the IEEE.

References

[1] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.

[2] F. Weng, P. Angkititrakul, E. E. Shriberg, L. Heck, S. Peters, and J. H. L. Hansen, "Conversational in-vehicle dialog systems: The past, present, and future," *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 49–60, Nov. 2016.

[3] B. W. Schuller, "The computational paralinguistics challenge," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 97–101, July 2012.

[4] C. Moseley, *Atlas of the World's Languages in Danger*, 3rd ed. Paris: Unesco Publishing, 2010.

[5] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.

[6] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Process.*, vol. 7, no. 3–4, pp. 197–387, June 2014.

[7] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. Int. Conf. Machine Learning (ICML)*, New York, 2016, pp. 173–182.

[8] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[10] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.

[11] M. Harper. IARPA Babel program. Intelligence advanced research projects activity, Office of the Director of National Intelligence, Washington, D.C. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/babel

[12] B. W. Schuller, "Speech analysis in the big data era," in *Text, Speech, and Dialogue* (Lecture Notes in Computer Science, vol. 9302), P. Král and V. Matoušek, Eds. Berlin: Springer-Verlag, 2015, pp. 3–11.

[13] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3169–3173.

[14] M. R. Robertson. (2015, Nov. 13). 500 hours of video uploaded to YouTube every minute. *Tubular Insights*. [Online]. Available: http://www.reelseo.com/hours-minute-uploaded-youtube

[15] M. Eskénazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Hoboken, NJ: Wiley, 2013.

[16] J. D. Williams, I. D. Melamed, T. Alonso, B. Hollister, and J. Wilpon, "Crowdsourcing for difficult transcription of speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, 2011, pp. 535–540.

[17] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[18] Z. Zhang, *Semi-Autonomous Data Enrichment and Optimisation for Intelligent Speech Analysis*. Munich, Germany: Verlag Dr. Hut, 2015.

[19] A. Nagórski, L. Boves, and H. J. Steeneken, "Optimal selection of speech data for automatic speech recognition systems," in *Proc. INTERSPEECH*, Denver, CO, 2002, pp. 2473–2476.

[20] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proc. Asia-Pacific Signal and Information Processing Assoc. Annu. Summit and Conf. (APSIPA)*, Hong Kong, China, 2015, pp. 1225–1237.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[22] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[23] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP)*, Honolulu, HI, 2008, pp. 254–263.

[24] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Proc. Human Language Technologies: 2010 Annu. Conf. North American Chapter Assoc. Computational Linguistics*, Los Angeles, 2010, pp. 207–215.

[25] S. Hantke, T. Appel, F. Eyben, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 891–897.

[26] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, et al. "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8111–8115.

[27] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Berkeley, CA, 2010, pp. 312–317.

[28] A. Tarasov, S. J. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *Proc. W3C workshop on Emotion Markup Language (EmotionML)*, Paris, 2010, pp. 1–5.

[29] J. Ledlie, B. Odero, E. Minkov, I. Kiss, and J. Polifroni, "Crowd translator: On building localized speech recognizers through micropayments," *ACM SIGOPS Operating Syst. Rev.*, vol. 43, no. 4, pp. 84–89, Jan. 2010.

[30] C-Y. Lee and J. R. Glass, "A transcription task for crowdsourcing with automatic quality control," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 3041–3044.

[31] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 186–197, Jan. 2008.

[32] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 410–415.

[33] H. Wang, T. Lee, C. C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8545–8549.

[34] G. Mantena and X. Anguera, "Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8515–8519.

[35] X. Anguera, "Method and system for improved pattern matching," EP Patent EP12 382 508, 2012.

[36] Y. Chung, C. Wu, C. Shen, H. Lee, and L. Lee, "Audio Word2Vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 765–769.

[37] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 669–679, Apr. 2016.

[38] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 4366–4369.

[39] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 5532–5536.

[40] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, Atlanta, GA, 2013.

[41] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 23, no. 9, pp. 1469–1477, Sept. 2015.

[42] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 1420–1424.

[43] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3586–3589.

[44] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 3214–3218.

[45] B. Milde and C. Biemann, "Using representation learning and out-of-domain data for a paralinguistic speech task," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 904–908.

[46] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human emotion," *Int. J. Speech Technol.*, vol. 15, no. 3, pp. 313–323, June 2012.

[47] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust ASR," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, Merano, Italy, 2009, pp. 205–210.

[48] Z. Ling, S. Kang, H. Zen, A. Senior, M. Schuster, X. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 35–52, May 2015.

[49] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jan. 2006.

[50] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[51] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[52] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *J. Mach. Learning Res.*, vol. 11, pp. 625–660, Mar. 2010.

[53] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Machine Learning (ICML)*, Helsinki, Finland, 2008, pp. 1096–1103.

[54] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A-r. Mohamed, and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1692–1695.

[55] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten ZIP code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[56] O. Abdel-Hamid, A-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[57] D. Hau and K. Chen, "Exploring hierarchical speech representations with a deep convolutional neural network," in *Proc. 11th U.K. Workshop on Computational Intelligence (UKCI)*, Manchester, U.K., 2011, pp. 37–42.

[58] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Advances Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2009, pp. 1096–1104.

[59] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. Int. Conf. Machine Learning (ICML)*, New York, 2009, pp. 609–616.

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

[61] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Machine Learning (ICML)*, Lille, France, 2015, pp. 843–852.

[62] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. Workshop Syntax, Semantics and Structure Statistical Translation (SSST)*, Doha, Qatar, 2014, pp. 103–111.

[63] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.

[64] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 437–440.

[65] A-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[66] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1695–1699.

[67] Y. Liu, T. Fu, Y. Fan, Y. Qian, and K. Yu, "Speaker verification with deep features," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, Beijing, 2014, pp. 747–753.

[68] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5688–5691.

[69] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Deep learning for emotional speech recognition," in *Pattern Recognition*, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. A. Olvera-Lopez, J. Salas-Rodríguez, and C. Y. Suen, Eds. MCPR 2014. *Lecture Notes in Computer Science*, vol. 8495. Berlin: Springer-Verlag, 2014, pp. 311–320.

[70] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2013, pp. 3687–3691.

[71] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2725–2728.

[72] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 23–31, Jan. 2005.

[73] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, "Speaker identification using semi-supervised learning," in *Proc. 17th Int. Conf. Speech and Computer (SPECOM)*, Athens, Greece, 2015, pp. 389–396.

[74] R.-C. Hsiao, T. Ng, F. Grézl, D. Karakos, S. Tsakalidis, L. Nguyen, and R. Schwartz, "Discriminative semi-supervised training for keyword search in low resource languages," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 440–445.

[75] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6704–6708.

[76] Z. Zhang, J. Deng, and B. Schuller, "Co-training succeeds in computational paralinguistics," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8505–8509.

[77] X. Cui, J. Huang, and J.-T. Chien, "Multi-view and multi-objective semi-supervised learning for HMM-based automatic speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1923–1935, Sept. 2012.

[78] Y. Liu and K. Kirchhoff, "Graph-based semisupervised learning for acoustic modeling in automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1946–1956, Nov. 2016.

[79] G. Riccardi and D. Hakkani-Tür, "Active learning: Theory and applications to automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 504–511, July 2005.

[80] B. Varadarajan, D. Yu, L. Deng, and A. Acero, "Maximizing global entropy reduction for active learning in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Taibei, China, 2009, pp. 4721–4724.

[81] T. Fraga-Silva, J.-L. Gauvain, L. Lamel, A. Laurent, V.-B. Le, and A. Messaoudi, "Active learning based data selection for limited resource STT and KWS," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 47–53.

[82] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, 2010, pp. 4350–4353.

[83] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. INTERSPEECH*, Portland, OR, 2012, pp. 362–365.

[84] Y. Zhang, E. Coutinho, Z. Zhang, M. Adam, and B. Schuller, "On rater reliability and agreement based dynamic active learning," in *Proc. 6th Biannu. Conf. Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 70–76.

[85] G. Riccardi and D. Z. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in *Proc. INTERSPEECH*, Geneva, Switzerland, 2003, pp. 1825–1828.

[86] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Commun.*, vol. 52, no. 7, pp. 652–663, Aug. 2010.

[87] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Comput. Speech & Language*, vol. 24, no. 3, pp. 433–444, July 2010.

[88] X. Zhu, "Semi-supervised learning literature survey," Department of Computer Sciences, University of Wisconsin, Madison, Tech. Rep. TR 1530, 2006.

[89] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2000.

[90] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, HI, 2011, pp. 523–528.

[91] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5185–5189.

[92] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Computational Learning Theory (COLT)*, Madison, WI, 1998, pp. 92–100.

[93] H. Xu, H. Su, C. Ni, X. Xiao, H. Huang, E. S. Chng, and H. Li, "Semi-supervised and cross-lingual knowledge transfer learnings for DNN hybrid acoustic models under low-resource conditions," in *Proc. INTERSPEECH*, San Francisco, 2016, pp. 1315–1319.

[94] B. Settles, "Active learning literature survey," Department of Computer Sciences, University of Wisconsin, Madison, Tech. Rep. TR 1648, 2009.

[95] N. Roy and A. McCallum, "Toward optimal active learning through Monte Carlo estimation of error reduction," in *Proc. 18th Int. Conf. Machine Learning (ICML)*, Williamstown, MA, 2001, pp. 441–448.

[96] J. Zhu, H. Wang, B. K. Tsou, and M. Ma, "Active learning with sampling by uncertainty and density for data annotations," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1323– 1331, Aug. 2010.

[97] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learning*, vol. 28, no. 2-3, pp. 133–168, Aug. 1997.

[98] A. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proc. Int. Conf. Machine Learning (ICML)*, Madison, WI, 1998, pp. 359–367.

[99] I. Muslea, S. Minton, and C. Knoblock, "Active + semi-supervised learning = robust multi-view learning," in *Proc. Int. Conf. Machine Learning (ICML)*, Sydney, Australia, 2002, pp. 435–442.

[100] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU)*, Kyoto, Japan, 2007, pp. 562–565.

[101] C. E. Erdem, E. Bozkurt, E. Erzin, and A. T. Erdem, "RANSAC-based training data selection for emotion recognition from spontaneous speech," in *Proc. 3rd Int. Workshop on Affective Interaction in Natural Environments (AFFINE).* New York, 2010, pp. 9–14.

[102] Z. Zhang, F. Eyben, J. Deng, and B. Schuller, "An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena," in *Proc. 5th Int. Workshop Emotion Social Signals, Sentiment & Linked Open Data (satellite of LREC 2014)*, Reykjavik, Iceland, 2014, pp. 21–26.

[103] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.

[104] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intell. Res.*, vol. 16, pp. 321–357, June 2002.

[105] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 312–315.

[106] A. I. García-Moral, R. Solera-Ureña, C. Peláez-Moreno, and F. Díaz-de María, "Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 468–481, Mar. 2011.

[107] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, Oct. 2000.

[108] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Proc. Advances Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2006, pp. 601–608.

[109] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007, pp. 1433–1440.

[110] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *J. Mach. Learning Res.*, vol. 10, pp. 1391–1445, July 2009.

[111] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1458–1468, July 2013.

[112] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sept. 2014.

[113] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2897–2901.

[114] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3311–3315.

[115] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.

[116] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, Geneva, Switzerland, 2013, pp. 511–516.

[117] A. Kocsor and L. Tóth, "Kernel-based feature extraction with a speech technology application," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2250–2263, Aug. 2004.

[118] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1025–1031, Sept. 2011.

[119] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8619–8623.

[120] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7304–7308.

[121] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 11, pp. 1938–1949, Nov. 2015.

[122] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 4818–4822.

[123] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5014–5018.

[124] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[125] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[126] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5800–5804.

[127] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, June 1996.

[128] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6965–6969.

[129] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[130] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Machine Learning (ICML)*, Bellevue, WA, 2011, pp. 689–696.

[131] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001, pp. 11–19.

[132] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Commun.*, vol. 75, pp. 27–49, Dec. 2015.

[133] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. I–237–I–240.

[134] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affective Comput.*, vol. 1, no. 2, pp. 119–131, July 2010.

[135] S. Y. Kung, "Compressive privacy: From information/estimation theory to machine learning," *IEEE Signal Process. Mag.*, vol. 34, no. 1, pp. 94–112, Jan. 2017.

[136] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Trans. Affective Comput.*, vol. 5, no. 4, pp. 406–417, Oct. 2014.

[137] R. Fierimonte, S. Scardapane, A. Uncini, and M. Panella, "Fully decentralized semi-supervised learning via privacy-preserving matrix completion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–13, 2016.

[138] S. Huffman. (2014, Oct. 14). OMG! Mobile voice survey reveals teens love to talk. [Online]. Available: https://googleblog.blogspot.de/2014/10/omg-mobile-voice-survey-reveals-teens.html

[139] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha, and L. Zhao, "Practical speech emotion recognition based on online learning: From acted data to elicited data," *Math. Problems in Eng.*, vol. 2013, pp. 9, June 2013.

[140] W. Ainsworth and S. Pratt, "Feedback strategies for error correction in speech recognition systems," *Int. J. Man-Mach. Stud.*, vol. 36, no. 6, pp. 833–842, June 1992.

[141] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, et al. "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.