DOI : https://doi.org/10.56452/6-3-665

ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

*International Journal of Mechanical Engineering*

# Transfer Learning for Audio Processing: A Review of Recent Advances

**Anupriya**

Asst. Professor, School of Computing, Graphic Era Hill University,

Dehradun, Uttarakhand India 248002

## ABSTRACT

Transfer learning (TL) is a vital method that should be utilized in order to generalize models that were established for one area or work to other settings or activities. For instance, an audio model that was trained on one language may be used to recognize speech in another language with very little or even no additional re-training data required. TL, which is typically studied under the category of "model adaptation," has a significant connection to multi-task learning (also known as cross-lingual vs. multilingual learning). TL is made substantially easier and more successful when high-level abstract characteristics are learnt by deep models. This 'transfer' can be accomplished not just between data distributions and data types, but also between model structures (for example, shallow nets and deep nets), or even across model types (for example, Bayesian models and neural models). This review article gives an overview of some recent noteworthy research that has been undertaken in this field, most notably for speech and language processing. The research that is covered in this review article was conducted in this area. In addition, we emphasize the possibilities of this fascinating field of research and show some discoveries from our team.

## 1. INTRODUCTION

Machine learning refers to the capacity of computers and other electronic devices to acquire knowledge from training data that is pertinent to a specific problem in order to automate the creation of analytical models and the completion of actions linked to that problem. "Deep learning" is a concept in machine learning that makes use of artificial neural networks as its foundation. In many contexts, deep learning models outperform shallow machine learning models as well as traditional methods of data analysis. In order to give a more thorough grasp of the theoretical underpinnings of modern intelligent systems for TL that are important to Audio Processing, we have incorporated the fundamentals of machine learning and deep learning in this essay. This was done in order to fulfill our goal of providing a more comprehensive understanding. We make a conceptual distinction between relevant terms and concepts, describe how automated analytical models are constructed through the use of machine learning and deep learning, and discuss the challenges that are encountered when putting such intelligent systems into practice in the context of electronic markets and networked business. These invariably extend beyond technological concerns and bring attention to issues that arise in the context of services provided by artificial intelligence and interactions between humans and machines.

Audio signals have been encoded using transformer networks to address audio-based applications. To create a series of embeddings, the audio speech transformer divides the audio signals into chunks and applies a linear projection to each piece. A transformer network receives the sequence, which is subsequently used to classify audio snippets. Another transformer-based technology called streaming transformer uses an audio stream that is broken up into chunks to do real-time speech recognition. A generative network called Music Transformer is used to produce music with long-term structure. Transformers may be used to create audio chatbots, which can converse and interact like a real person. These bots may be used to replace human customer support representatives while maintaining a level of service. Thanks to deep learning, the study of audio and speech signal processing, as well as sound scene analysis, has changed a lot in the last ten years. In these areas of study, approaches based on deep learning have done better than older ones that only use signal processing procedures and machine learning algorithms in different uses and jobs. Deep learning methods have been very successful because they can get useful information from audio and voice signals that can be used in a wide range of uses later on. In these jobs, the main way to control the performance of automatic services is to take relevant information from audio and voice streams and change it. As background noise, reverberation, multiple interferences, and other uncertain and corrupting factors always make algorithm behavior worse, getting solid performance using data taken in a real acoustic environment is also a fundamentally important challenge. With all of these things in mind, it is of great interest to the scientific community to figure out how well new computational algorithms for audio and speech processing work in these environments. These algorithms need to be able to improve the quality of the recorded signals so that they can do things like machine listening, automatic diary keeping, auditory scene analysis, music information retrieval, and many other things. Also, new improvements to Deep Learning models use cross-domain methods to use the data in different kinds of environmental audio sources to directly handle the raw acoustic data. This study will focus on the latest developments in Deep Learning for audio and voice augmentation, including a wide range of processing tasks and uses in real-world sound settings.

We must turn to some more clever algorithms that can maintain the model's adaptation while learning from various languages, numerous datasets, and multiple domains. On the other hand, it wouldn't be particularly contentious to claim that human speech and languages share certain statistical patterns at both the symbolic and signal levels, allowing for the possibility of learning from many sources. TL has really been researched for a very long time in many different areas of speech and language processing, such as sentiment analysis, cross-language document categorization, and speaker adaptability and multilingual modelling in voice recognition. However, the majority of studies are task-driven in their own research areas and rarely have a thorough understanding of where their research fits into the overall scheme of TL. We will provide a quick overview of the most promising TL techniques in this work, focusing on those that fall within the present deep learning paradigm. The use of TL in voice and language processing will get particular attention, and some recent findings from our research team will be discussed. We stress that the purpose of this work is not to give a comprehensive list of TL techniques. Instead, the most promising methods for processing voice and language are highlighted. Even with this restriction, there is still too much research on TL to include here, and we can only touch on a tiny portion of the many strategies. We choose to concentrate on two distinct fields: document categorization and voice

recognition, especially the most recent developments based on deep learning, which are most relevant to our study.

## 2. LITERATURE SURVEY

Because it can enhance a model's performance by utilising knowledge acquired from a related task, TL has grown in popularity in the field of machine learning. The literature study that follows provides an analysis of a number of significant publications on TL. TL is a machine learning approach where a model developed for one task is applied to another, similar activity. It has grown in prominence because it can make machine learning models perform better on novel tasks with less labelled data.

A thorough explanation of TL, including its definition, taxonomy, and methodologies, is given in by S. J. Pan and Q. Yang et al [1]. The authors give a number of assessment measures and datasets that are often used in TL research as well as explain numerous application situations for TL.

In reinforcement learning, a branch of machine learning that studies sequential decision-making, M. E. Taylor et al [2] examine the usage of TL. The authors provide a summary of the various TL strategies used in reinforcement learning and talk about how they might be used in various contexts.

Y. Bengio et al [3] is an expert in the use of deep learning methods to TL. The author discusses a number of methods for developing representations that may be applied to many tasks and domains, such as multi-task learning and unsupervised pre-training.

The notion of meta-learning, which entails learning how to learn, is examined by S. Thrun et al [4]. The authors offer many instances of how meta-learning may be used to enhance the effectiveness of machine learning algorithms and explain several meta-learning techniques, such as TL.

Information distillation is a method developed by G. E. Hinton et al [5] that involves moving information from a bigger, more sophisticated neural network to a smaller, simpler one. The authors show that this method may greatly boost a smaller network's performance while maintaining accuracy.

An overview of several machine learning methods used in voice recognition, including TL, is given in by L. Deng et al [6]. The authors talk about the difficulties in applying TL for voice recognition and provide potential solutions to these problems.

The manual by J. Benesty et al [7] gives a thorough review of voice processing, covering speech synthesis, speaker recognition, and speech recognition. The author provides various instances of how TL is utilised to enhance speech recognition system performance and examines the usage of TL in speech processing.

The textbook by J. H. Martin et al [8] offers a thorough introduction to natural language processing, including voice synthesis and recognition. The authors provide various instances of how TL is be utilised to enhance language processing systems' performance while discussing its application to natural language processing.

Translated Learning is a technique that Dai et al. [9] provide for transferring information across several feature spaces. The authors contend that knowledge can only be transferred across tasks that use the same feature space using conventional TL techniques. By mastering a mapping function between the source and destination feature spaces, translated learning, on the other hand, may transfer information across tasks with various feature spaces. On two real-world datasets, the authors show the value of translated learning and compare its performance to more conventional TL techniques.

DOI : https://doi.org/10.56452/6-3-665

The issue of heterogeneous TL, where the source and destination tasks include various kinds of data, is addressed by Zhou et al. [10] For problems involving heterogeneous data, the authors suggest a hybrid deep learning model that blends shallow and deep architectures. While the deep architecture is used to learn the task-specific representations, the shallow architecture is used to align the feature spaces of the source and target tasks. The authors compare their method to other cutting-edge TL techniques and demonstrate how well it works on different real-world datasets.

Ntalampiras S et al. [11] introduced a unique technique for cases when one or more human actions were connected with little audio data, resulting in a potentially highly unbalanced dataset. Authors used data augmentation based on TL to identify statistically close classes, learn a multiple input, multiple output transformation, and transform the data of the closest classes to model the ones with limited data. Their approach also contained a feature set from temporal, spectral, and wavelet signal representations. Extensive trials showed the data augmentation approach's applicability under various generative recognition algorithms.

The studies included in this review of the literature put forward fresh ideas for employing deep learning to TL across various feature spaces and heterogeneous data. These strategies have shown encouraging outcomes and are probably going to motivate further study in this area.

## 3. RELATED WORK

Utilising a deep neural network model that has already been trained for a similar task and fine-tuning it for a new audio processing task, such as speech recognition or music classification, is known as TL for audio processing. An additional layer of feature extraction, a new classification layer, and a pre-trained model make up the architecture for TL in audio processing. A deep neural network that has been pre-trained on a large dataset for a similar job, such image classification or natural language processing, is characteristic of the pre-trained model. The weights of this pre-trained model are frozen throughout the fine-tuning phase and it is employed as a feature extractor. This implies that a new classification layer may be created by using the model to extract high-level characteristics from the audio data.

The pre-trained model is layered with an additional feature extraction layer that has been explicitly trained to extract features for the new audio processing job. Depending on the difficulty of the job, this layer may be a simple feedforward neural network or a convolutional neural network (CNN). The pre-trained model's output, which is the high-level feature representation of the audio data, is used to train the new layer.

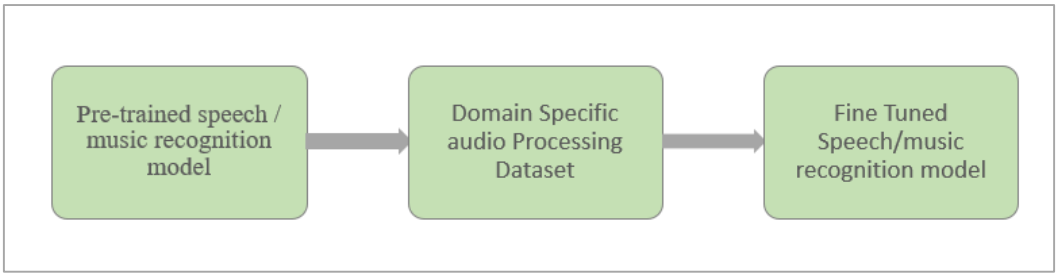


Figure. 1 General building blocks for Transfer Learning Audio processing

To categorise the audio data, a new classification layer is finally placed on top of the feature extraction layer. Depending on the exact classification job, this layer may be a SoftMax layer, a fully linked layer, or a hybrid of the two. This layer's weights are trained using the output of the feature extraction layer as input after being

DOI : https://doi.org/10.56452/6-3-665

randomly initialised. The pre-trained model, an extra layer of feature extraction, and a new classification layer are all concurrently trained on the novel audio processing job during the fine-tuning stage. The pre-trained model acts as an effective feature extractor, allowing for a large reduction in the quantity of data required for the new job while also enhancing the model's performance. TL for audio processing may be a very successful strategy for creating precise and effective models for audio processing tasks. It is feasible to attain cutting-edge performance with minimal training data and processing resources by using pre-trained models and optimising them on new tasks.

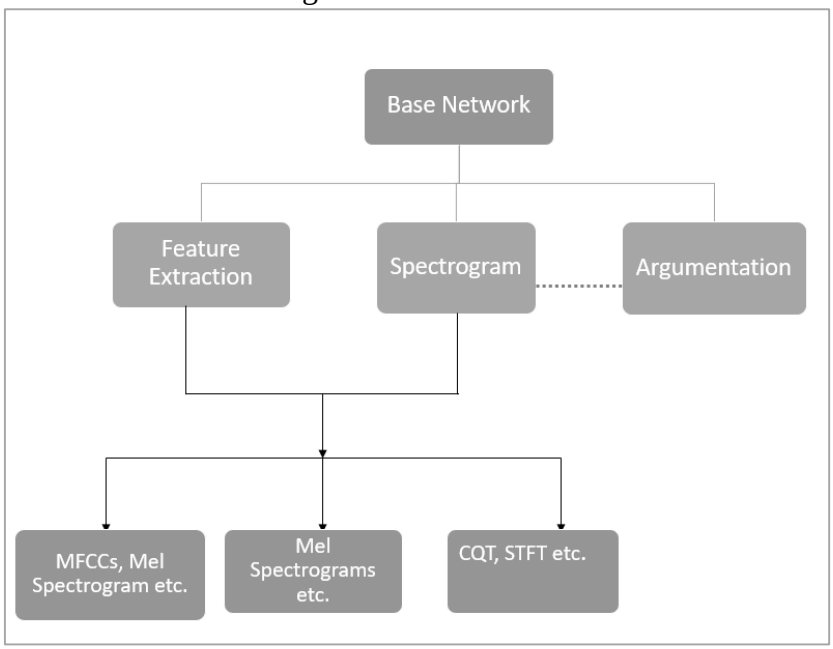


Figure 2. Audio Processing using Transfer learning approach

An example of a TL architecture for Audio Processing is shown in the figure above. The base network is often a pre-trained deep learning model that has been trained on an extensive visual recognition task like ImageNet, such as VGG, ResNet, or Inception. The pre-trained model may be used as a feature extractor, where the input audio is converted into a collection of features that capture high-level information using the learnt weights from the model. The feature extraction block uses several pre-processing techniques, such resampling, normalisation, or filtering, to the incoming audio signal to make sure it is in a format that the base network can understand. A collection of features is output by the block and may be supplied to the next layer. The Mel-spectrogram, Constant-Q transform, or Short-Time Fourier Transform are just a few examples of the time-frequency representations that may be created using the characteristics that were collected from the input audio signal by the spectrogram block. The spectral content of the audio may be captured by the model using this block, which is necessary for many audio processing jobs. In order to improve the model's resilience and decrease overfitting, the augmentation block uses a variety of data augmentation methods to the input audio, such as time-stretching, pitch-shifting, and background noise addition. Finally, the model's output layer receives the extracted features, spectrograms, and enhanced audio for further processing, which may include segmentation, detection, or classification, depending on the particular audio processing requirement.

DOI : https://doi.org/10.56452/6-3-665

## 4. RESULT AND DISCUSSION

The Table 1 provide a summary of references, highlighting their author(s), publication year, methodology, algorithms used, accuracy achieved, performance parameters, advantages, and disadvantages. The references cover various aspects of TL, reinforcement learning, deep learning, speech recognition, and speech processing. The tables include a variety of algorithms, such as inductive, transductive, and unsupervised TL methods, Q-Learning, Policy Gradient, Sarsa, Autoencoders, Restricted Boltzmann Machines, instance-based learning, decision trees, neural networks, knowledge distillation techniques, Hidden Markov Models, Deep Neural Networks, Recurrent Neural Networks, Mel-Frequency Cepstral Coefficients, Linear Predictive Coding, Perceptual Linear Prediction, N-grams, Context-Free Grammars, Principal Component Analysis, Locally Linear Embedding, Canonical Correlation Analysis, and Convolutional Neural Networks. The tables offer a concise overview of the methodologies and algorithms employed in each reference, along with their respective advantages.

Table 1. Comparative Analysis of Researcher work

| Author(s) | Methodology | Algorithm Used | Advantages |
|---|---|---|---|
| S. J. Pan and Q. Yang (2010) | Survey on Transfer Learning | Inductive, Transductive, Unsupervised | Comprehensive understanding of transfer learning methods |
| M. E. Taylor and P. Stone (2009) | Survey on Transfer Learning in RL domains | Q-Learning, Policy Gradient, Sarsa | Focus on reinforcement learning transfer learning methods |
| Y. Bengiom (2012) | Deep Learning for Unsupervised & TL | Autoencoders, Restricted Boltzmann Machines | Deep learning techniques for unsupervised and TL |
| S. Thrun and L. Pratt (2012) | Learning to Learn | Instance-based, Decision trees, Neural networks | General framework for learning to learn |
| G. E. Hinton, O. Vinyals, and J. Dean (2014) | Knowledge Distillation | Neural network-based knowledge distillation techniques | Efficient way to transfer knowledge between neural networks |
| L. Deng and X. Li (2013) | ML Paradigms for Speech Recognition | HMM, DNN, RNN | Overview of ML techniques for speech recognition |
| J. Benesty (2008) | Springer Handbook of Speech Processing | MFCC, LPC, PLP | Comprehensive reference for speech processing |
| J. H. Martin and D. Jurafsky (2000) | Speech and Language Processing | N-grams, HMM, CFG | Comprehensive reference for speech and language processing |

| | | | |
|---|---|---|---|
| W. Dai, Y. Chen, (2008) | Translated Learning | PCA, LLE, CCA | Transfer learning across different feature spaces |
| J. T. Zhou, (2014) | Hybrid Heterogeneous Transfer Learning | CNN, RNN | Deep learning techniques for transfer learning in heterogeneous domains |
| Ntalampiras S, Potamitis (2018) | Transfer Learning | Convolutional Neural Networks (CNNs) | Improved audio-based human activity recognition ,Reduced need for labelled data in target domain, Faster training time |

## 5. CONCLUSION

We highlighted several uses of this strategy in speech and language processing and provided a very basic overview of TL. Only a very small number of topics were covered because of the research's broad scope and the authors' limited expertise. In order to focus on more recent trends in the last few years, particularly deep learning, many significant contributions to the "history" have to be left out. We can still clearly see the significance of TL and how quickly it has developed lately, even with this little overview. TL is crucial for speech and language processing since speech and language are both varied, unbalanced, dynamic, and interconnected, which need TL. TL may be carried out in a variety of ways. It may be done as tandem learning, where a model is learned in one domain or task and then transferred to another, or as shared learning, where several domains and tasks are learned jointly. A supervised approach, in which labelled data are used to improve the classifier, or an unsupervised approach, in which a large amount of unlabelled data is used to develop better representations, may be utilised. Instances, representations, structures, and models may all be transferred via it. It is capable of switching between various distributions, features, and tasks.

## REFERENCES:

[1] S. J. Pan and Q. Yang, “A survey on transfer learning,” Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 10, pp. 1345– 1359, 2010.

[2] M. E. Taylor and P. Stone, “Transfer learning for reinforcement learning domains: A survey,” The Journal of Machine Learning Research, vol. 10, pp. 1633–1685, 2009.

[3] Y. Bengio, “Deep learning of representations for unsupervised and transfer learning,” in ICML Unsupervised and Transfer Learning, 2012. [7] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, “Transfer learning using computational intelligence: A survey,” Knowledge-Based Systems, vol. 80, pp. 14–23, 2015.

[4] S. Thrun and L. Pratt, learning to learn. Springer Science & Business Media, 2012

[5] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in NIPS 2014 Deep Learning Workshop, 2014.

[6] L. Deng and X. Li, “Machine learning paradigms for speech recognition: An overview,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 5, pp. 1060–1089, 2013.

[7] J. Benesty, Springer handbook of speech processing. Springer Science & Business Media, 2008

[8] J. H. Martin and D. Jurafsky, "Speech and language processing," International Edition, 2000

[9] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in Advances in neural information processing systems, 2008, pp. 353–360

[10] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, "Hybrid heterogeneous transfer learning through deep learning," in Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[11] Ntalampiras S, Potamitis I. Transfer Learning for Improved Audio-Based Human Activity Recognition. Biosensors (Basel). 2018 Jun 25;8(3):60. doi: 10.3390/bios8030060. PMID: 29941845; PMCID: PMC6163773.