**CHAPTER 2 LITERATURE SURVEY**

The major goal of the application is audio classification based on SNR and rate of speech. This requires references on various inherited and self-possessed properties on analog signals and their representation. The sample analog audio is converted into digital format in order to extract and understand the features of the audio sample for classification. On this digital format basis, the well-suited audio characteristic feature has to be chosen in order to build a consolidated application for the classification. These below references are used for this survey to equalize the suitable digital format and the audio feature to be used in building the model.

|  |  |  |
| --- | --- | --- |
| **TITLE** | **AUTHOR** | **YEAR OF PUBLICATION** |
| Multiengine Speech | Ahmad R. Abu-El- | February 2012 |
| Processing Using SNR | Quran, Adrian D. C. |  |
| Estimator in Variable | Chan, and Rafik A. |  |
| Noisy Environments | Goubran |  |
| Optimizing spectral | R. Gomez and T. | March 2010. |
| subtraction and Wiener | Kawahara, in IEEE |  |
| filtering for robust | International |  |
| speech recognition in | Conference on |  |
| reverberant and noisy | Acoustics, Speech, and |  |
| conditions | Signal Processing |  |

|  |  |  |
| --- | --- | --- |
| Audio-visual speech recognition using deep learning | [Kuniaki Noda](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-1), [Yuki Yamaguchi](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-2), [Kazuhiro Nakadai](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-3), [Hiroshi G. Okuno](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-4) , [Tetsuya Ogata](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-5) | December 2014 |
| The Short-Time Silence of Speech Signal as Signal-To-Noise Ratio  Estimator | Azhar S. Abdulaziz1, Veton Z. Këpuska | August 2016 |
| A Closer Look at Weak Label Learning for Audio Events | Ankit Shah Anurag Kumar Alexander G.Hauptmann  Bhiksha Raj | April 2018 |
| Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals | Sören Becker Marcel Ackermann  Sebastian Lapuschkin Klaus-Robert Müller Wojciech Samek | July 2018 |
| Introduction of the speaking rate in the model of speech recognition | Abdellah Yousfi Abdelouafi Meziane | February 2010 |
| pyAudioAnalysis: An Open-Source Python Library for Audio  Signal Analysis | Theodoros Giannakopoulos | December 2015 |

|  |  |  |
| --- | --- | --- |
| **REFERENCES** | | |
| Deep Learning With Python:  Machine learning mastery | Jason Brownlee | 2017 |
| Unsupervised feature learning for audio classification using convolutional deep  belief networks | Honglak Lee Yan Largman Peter Pham Andrew Y. Ng | 2009 |

**Table 2.1: Literature Survey**

1. **“MULTIENGINE SPEECH PROCESSING USING SNR ESTIMATOR IN VARIABLE NOISY ENVIRONMENTS ”**

**AHMAD R. ABU-EL-QURAN, ADRIAN D. C. CHAN, AND RAFIK A. GOUBRAN**

**FEBRUARY 2012**

This approach uses a multiengine speech processing system that can detect the location and the type of audio signal in variable noisy environments. This system detects the location of the audio source using a microphone array; the system examines the audio first, determines if it is speech/non-speech, then estimates the value of the signal to noise () using a Discrete-Valued SNR Estimator. Using this SNR value, instead of trying to adapt the speech signal to the speech processing system, we adapt the speech processing system to the surrounding environment of the captured speech signal. In this paper, we introduced the Discrete-Valued SNR Estimator and a multiengine classifier, using Multiengine Selection or Multiengine Weighted Fusion. Also, we use the SI as example of the speech processing. The Discrete-Valued SNR Estimator achieves an accuracy of 98.4% in characterizing the environment. Compared to a conventional single engine SI system, the improvement in accuracy was as high as 9.0% and 10.0% for the Multiengine Selection and Multiengine Weighted Fusion, respectively.

1. **“AUDIO-VISUAL SPEECH RECOGNITION USING DEEP LEARNING”**

[**KUNIAKI NODA**](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-1)**,** [**YUKI YAMAGUCHI**](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-2)**,** [**KAZUHIRO NAKADAI**](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-3)**,** [**HIROSHI G. OKUNO**](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-4) **,**[**TETSUYA OGATA**](https://link.springer.com/article/10.1007/s10489-014-0629-7#auth-5)

**DECEMBER 2014**

1. **“PYAUDIOANALYSIS: AN OPEN-SOURCE PYTHON LIBRARY FOR AUDIO SIGNAL ANALYSIS”**

**THEODOROS GIANNAKOPOULOS DECEMBER 2015**

Audio-visual speech recognition (AVSR) system is thought to be one of the most promising solutions for reliable speech recognition, particularly when the audio is corrupted by noise. However, cautious selection of sensory features is crucial for attaining high recognition performance. In the machine-learning community, deep learning approaches have recently attracted increasing attention because deep neural networks can effectively extract robust latent features that enable various recognition algorithms to demonstrate revolutionary generalization capabilities under diverse application conditions. This study introduces a connectionist-hidden Markov model (HMM) system for noise-robust AVSR. By preparing the training data for the network with pairs of consecutive multiple steps of deteriorated audio features and the corresponding clean features, the network is trained to output de-noised audio features from the corresponding features deteriorated by noise.

Second, a convolutional neural network (CNN) is utilized to extract visual features from raw mouth area images. By preparing the training data for the CNN as pairs of raw images and the corresponding phoneme label outputs, the network is trained to predict phoneme labels from the corresponding mouth area input images. Finally, a multi-stream HMM (MSHMM) is applied for integrating the acquired audio and visual HMMs independently trained with the respective features. By comparing the cases when normal and de-noised Mel-frequency cepstral coefficients (MFCCs) are utilized as audio features to the HMM, our unimodal isolated word recognition results demonstrate that approximately 65 % word recognition rate gain is attained with de-noised MFCCs under 10 dB signal- to-noise-ratio (SNR) for the audio signal input. Moreover, our multimodal isolated word recognition results utilizing MSHMM with de-noised MFCCs and

Literature Survey Chapter 2

acquired visual features demonstrate that an additional word recognition rate gain is attained for the SNR conditions below 10 dB.

1. **“UNSUPERVISED FEATURE LEARNING FOR AUDIO CLASSIFICATION USING CONVOLUTIONAL DEEP BELIEF NETWORKS”**

**HONGLAK LEE, YAN LARGMAN, PETER PHAM, ANDREW Y. NG 2009**

1. **“A CLOSER LOOK AT WEAK LABEL LEARNING FOR AUDIO EVENTS”**

**ANKIT SHAH, ANURAG KUMAR, ALEXANDER G.HAUPTMANN, BHIKSHA RAJ**

**APRIL 2018**

Based on audio analysis the most preferable digital format should portray variations between different audio samples. A mono-channel audio is usually preferred for a clean classification over the multi-channel audio formats. A 16-bit PCM (pulse code modulation) is used to convert analog signals to digital formats. Pulse-code modulation (PCM) is a method used to digitally represent sampled analog signals. It is the standard form of digital audio in computers, compact discs, digital telephony and other digital audio applications. In a PCM stream, the amplitude of the analog signal is sampled regularly at uniform intervals, and each sample is quantized to the nearest value within a range of digital steps.

1. **“OPTIMIZING SPECTRAL SUBTRACTION AND WIENER FILTERING FOR ROBUST SPEECH RECOGNITION IN REVERBERANT AND NOISY CONDITIONS”**

**R. GOMEZ AND T. KAWAHARA, IN IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING**

**MARCH 2010**

Literature Survey Chapter 2

1. **“THE SHORT-TIME SILENCE OF SPEECH SIGNAL AS SIGNAL- TO-NOISE RATIO ESTIMATOR”**

**AZHAR S. ABDULAZIZ1, VETON Z. KËPUSKA AUGUST 2016**

The digitized audio can vary depending on the application aspects. This variation has effect on various audio features. The study on different features proves that most of the audio variation is categorized by MFCC feature which is represented widely using 40 coefficients of variation. In sound processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of- a-spectrum"). The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFCC, the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

1. **“INTRODUCTION OF THE SPEAKING RATE IN THE MODEL OF SPEECH RECOGNITION”**

**ABDELLAH YOUSFI, ABDELOUAFI MEZIANE FEBRUARY 2010**

The variation can be best suited for audio signal processing in a natural environment. The speech classification needs other qualities of an audio to assist the difference. Based on references on speech recognition, the spectral contrast features for speech analysis. Spectral contrast considers the spectral peak, the spectral valley, and their difference in each frequency sub-band.

1. **“INTERPRETING AND EXPLAINING DEEP NEURAL NETWORKS FOR CLASSIFICATION OF AUDIO SIGNALS”**

Literature Survey Chapter 2

**SÖREN BECKER, MARCEL ACKERMANN, SEBASTIAN LAPUSCHKIN, KLAUS-ROBERT MÜLLER, WOJCIECH SAMEK JULY 2018**

1. **“DEEP LEARNING WITH PYTHON:MACHINE LEARNING MASTERY”**

**JASON BROWNLEE 2019**

On the basis of the input data the most preferable method of speech recognition and audio classification show higher efficiency on neural networks. On the process of understanding the data and analyzing the features of it the best suitable type of neural network could be incorporated. The deep neural network would have best effect on prediction and classification for the application.

Hence the above studies propose an evident method in for rendering a model which inherits the classification objectives of the application with maximum accuracy rate.